DOCUMENT RESUME

ED 205 547 TM 810 426

AUTHOR Smith, Richard M.
TITLE Person Fit Analysis with the Rasch Model.
PUB DATE 1 Apr 81
NOTE 140p.

EDRS PRICE MF01/PC06 Plus Postage.
DESCRIPTORS Algebra: Error of Measurement: *Goodness of Fit: Guessing (Tests): High Schools: *Knowledge Level: *Latent Trait Theory: Mathematical Formulas: *Mathematical Models
IDENTIFIERS New Jersey College Basic Skills Placement Test: *Rasch Model

ABSTRACT

In all latent trait models there is an equation which describes the probabilistic outcome of the interaction of a person with a known ability and an item with a known difficulty. This modeled outcome can be compared to the observed outcome, and the residuals can be summed over the items to assess the fit of the person to the model. Various approaches to this "person fit" analysis are explored, and the mathematics of person fit analysis with the Rasch model is briefly discussed. Various patterns of responses which seem to provide meaningful information about the individuals taking the test are reviewed. They include guessing, startup (fumbling), plodding, content interaction, sloppiness (carelessness), reanalysis, and random responses. Data from a specific algebra test are used to illustrate the extent that these patterns actually occur. (Author/BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED205547

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
R. Smith

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

PERSON FIT ANALYSIS WITH THE RASCH MODEL

BEST COPY AVAILABLE

TM 810 426

Richard M. Smith
University of Chicago
April 1, 1981

Person Fit Analysis With The Rasch Model

I. Introduction

II. The History of Person Fit Analysis

III. The Mathematics of Person Fit Analysis

IV. Examples of Person Fit Analysis

- A. Guessing
- B. Startup (Fumbling)
- C. Plodding
- D. Content Interaction
- E. Sloppiness (Carelessness)
- F. Reanalysis
- G. Random Responses

V. A Study of The Frequency of Person Disturbances in Measurement

VI. Conclusion

I. INTRODUCTION

The idea of person fit analysis is a topic that has received considerable attention with the development on latent trait models in general and the Rasch model in particular. In all latent trait models there is an equation which describes the probabilistic outcome of the interaction of a person with a known ability and an item with a known difficulty. This probability or modeled outcome can be compared to the observed outcome of the interaction between the person and the item to create a residual. These residuals, which contain the magnitude of the departure from the predicted outcomes for every item-person interaction, can be summed over either the items or the people. Once summed these residuals can serve as the basis for assessing the fit of either the person or the item to the model.

This type of analysis can be applied to people in a variety of ways by summing the residuals over various groupings of the items that the person interacted with on the test to answer the question "How well does a person's test score represent his true ability?". We will first explore the various approaches that others have used in attempting to answer this question, and highlight the aspects of these approaches which have led us to consider the method proposed here. We will

continue with a brief discussion of the mathematics of person fit analysis with the Rasch model. Finally we will review various patterns of responses which seem to provide meaningful information about the individuals taking the tests and look at a specific elementary algebra test to illustrate the extent that the patterns discussed occur in real data.

II. THE HISTORY OF PRRSON FIT ANALYSIS

The idea of "person fit" is one of the central notions of the measurement process. If test scores are to have any relevance in estimating the ability of the individual we must first be able to say that the vector of responses to the items which is used to generate the score meets the standards we hold for the measurement process. This notion seems to be fundamental to Lumsden's (1976) conceptulization of the central questions of test theory. "How can we best use the information given us by the item performances of our subjects? How well do the test scores represent the attributes?"

Whether we recognize it or not, we all have a standard of measurement. That standard is that a test should give an ability estimate such that performance on an individual item or set of items is consistent with ability. This standard is often unexpressed or unrecognized but one clearly exists. If we were to look at the responses to a 10 item test in which the items were ordered from the easiest to the most difficult, we could begin to get a grasp on this standard. (This is an elaboration of an arguement developed in Wright and Stone (1979)).

	Easy			Average				Hard			
	1	2	3	4	5	6	7	8	9	10	SCORE
A	1	1	1	1	0	1	0	0	0	0	5
B	1	1	1	0	0	1	0	1	0	0	5
C	1	0	1	0	1	0	1	0	1	0	5
D	0	0	0	0	1	0	1	1	1	1	5

Clearly, if the notion of easy and hard items are to have any meaning, we must expect that individuals taking the test will generally be able to answer the easy items correctly more frequently and be able to answer the hard items correctly less frequently. Any situation which contradicts this pattern goes against the intuitive understanding of measurement that we all hold.

In the response vectors shown above all of the individuals have a "score" of 5. Individual A, though, has the most believable response vector. It is almost too believable. This individual has answered all of the easy items correctly, missed all of the hard items, and answered half of the average difficulty items correctly. Contrast this with individual B. In this case all of the easy items are correct, but only 1 of the 4 average difficulty items are correct, and 1 of the 3 hard items is correct. We might wonder why this individual could answer one of the hard items correctly while being able to answer only one of the average items correctly. In most situations we would, most likely, accept this as a reasonable response vector, but it was unusual enough to make us stop and

think about it. The response vector for individual C begins to be more questionable. Why would an individual be able to answer the second hardest item correctly and miss the second easiest item. We might suspect that the individual was careless on the easy item, or that the hard item contained an isolated fact that the individual happened to know, but was uncharacteristic of his general knowledge. Even though there is a general progression from 67% of the easy items correct to 50% of the average items correct to 33% of the hard items correct, we would probably be curious enough about this pattern of responses to want to talk with the individual about the test. The response vector of individual D is the exact opposite of what we observed for individual A. This individual answered all of the hard items correctly, none of the easy items correctly, and half of the average items correctly.

I don't think that anyone could look at individual D's response vector and state that the score of 5 means the same thing for individual D as for A. The difference between the two is so great that we would probably be forced to take some kind of action, e.g., retest D, rescore the answer sheet to see if an error had been made the first time, ask other people about this person, look at the test itself to see if we were really testing two different skills, etc. Whatever we decide to do, the important thing is that at some point we began to question the way an individual responded to a set of items. The central

issue in any measurement situation is whether the ability estimate provides an accurate representation of the attribute values. The relationship of an individual's response vector to our idea of what should happen when an individual and an item interact is certainly an important part of that determination.

I hope, this little demonstration has convinced us that we really do have an intuitive grasp of what "person fit" is. Simply stated it is a measure of how believable a series of responses is. Had we looked only at the total score for these individuals we would have never questioned the results. The simple procedure of ordering the items from easiest to hardest and looking at how each person responded to the items gives us some insight into the problem, and it is a problem which underlies all psychological measurement. The fundamental question is, if it is such an obvious problem and so relatively easy to detect, why haven't we paid more attention to the problem long before this? Why indeed! A look at some of the attempts others have made to solve this problem might be helpful in answering this question.

Cronbach (1946) applied the work of Scherif and Cantril (1945) in the psychology of attitudes testing to the analyzing the lack of person fit by studying specific possible sources of error. The result was the notion of response sets. Cronbach defines a response set as "any tendency causing a person consistently to give different responses to test items than he

would when the same content is presented in a different form". He goes on to describe 6 different types of response sets which are applicable to Psychological and Educational tests. They are:

1) tendency to gamble
2) defination of judgement categories
3) inclusiveness
4) bias
5) speed vs. accuracy
6) miscellaneous response sets

Since we are limiting our discussion to multiple choice items with only one correct answer, only 3 of Cronbach's original notions appply. The tendency to gamble can be generally described as the problem of guessing. Cronbach notes that the observed variance of test scores of cautious guessers is restricted. In order to remove the effect of this response set he urges that all test instructions encourage the subjects to answer every item. The notion of bias used by Cronbach refers to the tendency to choose one response (A,B,C,D, etc.) over another. Cronbach hypothesized that less able individuals would only read the choices until they came to a plausible one and mark that choice. This would give the less able individuals a higher score on items with A and B as the correct answer and a lower score on items with C and D as the correct answer. Cronbach was unable to detect this type of response set in any multiple choice tests. The speed vs. accuracy response set refers to the phenomena of working quickly and carelessly or slowly and accurately. Speededness can produce misleading

results in speeded tests with rights-only scoring, while an overemphasis on accuracy can lower scores in any testing situation.

Cronbach (1950) carried the analysis of response sets even further. Using split-half reliability coefficients, he was able to determine that response sets are reliable, i.e., that they appear to be constant throughout the test. Cronbach explores two other consequences resulting from the presence of response sets in test data that are worth noting. In some instances, Cronbach argues, the presence of response sets can improve the reliability of the test and that they will almost always reduce the validity of the test. As Cronbach states; "tests are usually constructed to measure a trait defined by the content of the test items. If the form of the item permits response sets, two people having equal true scores on the content factor will often receive different scores on the test" thus reducing the validity of the test scores.

In general, Cronbach (1950) defines four types of response set variance:

1) chance variance
2) internally consistent variance which is constant for a given test but shifts over time
3A) stable response tendencies correlated with the test criterion
3B) stable response tendencies uncorrelated with the test criterion

Cronbach considers type 1 variance to be error variance, type 2 and 3B variance to be harmful variance, and type 3A as

potentially helpful variance. He notes, however, that it is impossible to seperate these 4 types of variance in any one test.

Most of Cronbach's suggestions for dealing with response sets involve the selection of items and item formats which discourage response sets and the development of clear, unambiguous instructions. However, he makes one interesting suggestion for individual corrections for response sets, the use of special scoring keys which weight the choices to counterbalance the response set. Finally in the extreme case Cronbach recommends invalidating the individual's score. "Eliminating cases with extreme control scores has the disadvantage of throwing out numerous subjects, but it is vastly better than treating the subjects as if the scores were valid."

Thurstone (1927) laid the foundation for an interesting approach to person fit analysis in the law of comparative judgement. Thurstone envisioned a stimulus dimension with the various stimuli ordered by intensity along this dimension. The individual could also be placed along this dimension by observing his responses to a variety of stimuli intensities. Those stimuli well within the individual's perception range will be consistently perceived correctly, those stimuli well beyond the individual's perception range will be consistently not perceived. The individual's threshold will then be the

point where the responses to the stimuli are most inconsistent. This type of system lends itself to a probabilistic interpretation where the probability of inconsistent responses will diminish as a function of the distance of the stimuli from the threshold of the individual.

The psychophysical law developed by Thurstone has applicability "not only to the comparison of physical stimulus intensities but also to qualitative comparative judgements". Thurstone's concept of "just noticeable difference is contingent on the fact that an observer is not consistent in his comparative judgements from one occasion to the next. He gives different comparative judgements on successive ocassions about the same pair of stimuli". Central to this concept is the fact that the inconsistency of responses is maximumized at the point of least noticeable difference.

Moiser (1940), following upon Thurstone's psychophysical model, argues that the underlying theorems in psychophysics and mental test theory can be reduced to a common base. From this common base he goes on to argue that the method of equal appearing intervals and the law of comparative judgement have a counterpart in mental test theory.

Moiser (1941) continued his use of the concepts of psychophysical measurement by developing a method for interpreting dichotomous (right, wrong) responses. In this context Mosier states that the following theorem for estimating

a person's ability.

> "When stimuli are arranged in order of stimulus magnitude along a quantitative continuum, but the response of any individual to any stimulus can take only one of two values then the unit frequency function of response as a function of stimulus magnitude will be the integral of the normal probability curve. Then also the characteristic response value of an individual (his test score) is that stimulus magnitude yeilding each of the two possible responses 50 per cent of the time, and the individual's variability is the standard deviation of the probability function thus integrated."

We have thus arrived at the point where, using Mosier's theorm, a person's ability and the item's difficulty are both assigned values along a single continuum and a person's ability estimate for a given test can be defined as the point of maximum inconsistency in responses to the items. We would expect that items which lie substantially below an individual's ability would be answered correctly a large part of the time. Conversely, we would expect that items which lie substantially above an individual's ability would be answered incorrectly a large part of the time. This model allows us to expect a certain pattern of correct and incorrect responses. Consistency of responses, either correct or incorrect, is, therefore, characteristic of responses to sets of items that lie considerably above or below a person's ability and is not a feature of measurement specifically targeted on an individual's ability.

Blaser (1949) is concerned with the "consistency of response" to test items. Consistency in this case, however,

refers to the characteristic of responding in the same way to the same item on different administrations, rather than consistency over a series of items, but the concepts are similar. The major concerns which Glaser voices are:

1) How can we interpret the test scores of inconsistent individuals?
2) Will an inconsistent individual exhibit inconsistency in other testing situations?
3) Is inconsistency related to personal or test administration factors?

Glaser's method of determining inconsistency was to count the number of items that an individual changed from correct to incorrect or from incorrect to correct on two different test administrations of the same items. Changes from one incorrect response to another incorrect response were not counted as inconsistent.

Glaser's solution to the inconsistency problem, which is based on Mosier's work, lies in the following statement. "In psychophysical measurement as the discrimination of a stimulus becomes more difficult, i.e., as the individual's threshold is approached from either direction on the scale of measurement, his responses become less consistent. The threshold is defined as the point of greatest inconsistency of response." He argues psychological test items can be thought of as psychophysical stimuli. In this regard, as items become closer to an individual's ability, the difficulty of making a response reaches a maximum, and the individual will exhibit the highest degree of inconsistency at this point. If a person responds

according to this model, we expect as the items become either much harder or much easier, the individual will answer either consistently incorrectly or consistently correctly.

Glaser(1950, 1951) proposes a method of testing based on Ellson's concept of multiple operation measurement. This method is based on the principles of psychophysical measurement, and involves the administration "of a series of items spaced along a scale on which the test is designed to measure". For each individual the measure is based on the number of correct responses. The frequency of inconsistent responses in this model would be a function of the probability of a correct response at the scale positions of the items involved. Glaser (1951) found that when items are ordered in difficulty order there is an approximately normal distribution of inconsistent responses in the zone of transition from plus to minus. He concluded that the ability score is an estimate of the point of maximum inconsistency for an individual.

From this model Glaser was able to develop two hypotheses: first, that a test score may be considered to be an estimate of the maximum point of the distribution of inconsistent responses; and second, that if a test is uni-dimensional then the ordered items exhibit a distribution of inconsistent responses around the ability score in the area of maximum inconsistency. Using this technique Glaser argues that it is possible to identify the items belonging to each

factor on a two-dimentional test and to be able to estimate an individual's performance on each dimension by identifying the point of maximum inconsistent responses for each factor.

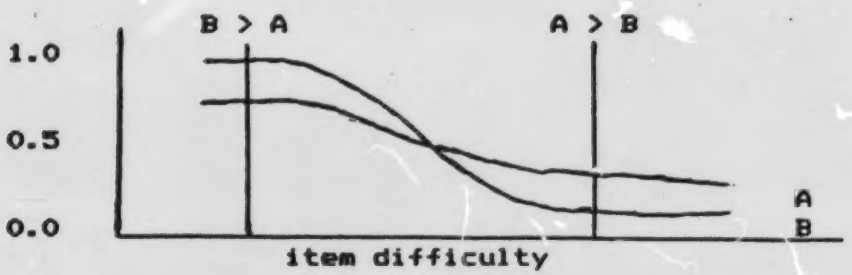
Lumsden (1977) considers the concept of person unreliability. Lumsden's purpose is "to present an attribute-based model of test performance and to derive from it a different, and potentially fruitful, approach to person reliability". Lumsden defines three types of change that can occur in a person's ability:

1) developmental trends: These are individual growth curves which represent long term trends.
2) swells: These are temporary short term fluctuations of ability lasting from several hours to days.
3) tremors: These are rapid momentary fluctuations in ability. These are thought of as random errors.

The concept of tremors is interesting. Cronbach in his discussions considered response sets to be constant over a single test. Thorndike (1949) did not consider changes in individuals during testing in his classification of the sources of variation in test scores. Lumsden contends that these within test changes play an important role in the interpretation of test scores. Using Weiss's (1973) notion of a "subject characteristic curve", a graph of a person's probability of responding correctly to items from a wide range of difficulties, to reflect the dispersion of the attribute locations resulting from tremors, Lumsden is then able to

derive the measure of person reliability from a comparison of the slopes of the person characteristic curves. It is not clear here whether Lumsden is proposing an additional person parameter or a measure of the person's fit to some modeled value. His later work, however, would tend to suggest that he would not endorse the concept of using the slope of the person characteristic curve as a person parameter.

In cases where the person characteristic curves for different people cross, the estimate of a person's ability as measured by the number of correct responses may be biased by the difficulty of the items making up the test. Lumsden plots the proportion of correct responses against the item difficulties to produce the person characteristic curves. Comparing two individuals with equal ability but different person characteristic curves we can see that an easy test would produce a higher ability estimate for person B and a difficult test would produce a higher estimated ability for person A.



The flatter person characteristic curve for person A is the result of a wider distribution of attribute locations resulting from stronger tremors than those experienced by person B.

Lumsden (1978) carries this argument further. Test reliability or "test score unreliability" as he terms it, is in fact the result of fluctuations in a person's ability, trends, swells and tremors, and not a property of tests or items. Again, his theory suggests that test items must have stable difficulties and, as a result, are perfectly reliable, hence, all tests are perfectly reliable. The unreliability of test scores results solely from person unreliability which can be linked directly to the tremors described above. The key to understanding this person unreliability lies in the person characteristic curve. The larger the amount of tremors the flatter the person characteristic curve will be. This theory has an interesting impact on the concept of sufficient statistics:

> "It follows generally that statements that particular estimation procedures yield sufficient statistics for estimating average ability levels for subjects may be misleading, even when the sufficiency claim is true there is another aspect of performance, the person reliability which cannot be estimated by a single test score."

Lumsden suggests that it is possible to define a reliability coefficient that will quantify the amount of person fluctuation. This statistic is "the proportion of variance in the obtained scores that is attributable to variance in the average attribute levels for subjects."

Lumsden (1980) notes that the person characteristic curves are required to have equal slopes in the Rasch model if

the specific objectivity of the ability estimates is to be preserved. As a result of this requirement he suggests a procedure also suggested by Andrich (1973), namely that all persons whose person characteristic curves did not conform to the average slope should be classified as unmeasurable on that test. The presence of unexpected item responses, which are thought of as indications of tremors, would be sufficient to invalidate an individual's test score if they appeared to alter the individual's person characteristic curve.

This discussion brings us full circle back to our original problem. We can see that the basic notion of measuring person fit comes from our expecting certain things to happen and attempting to describe the deviation of the actual occurrence from our expectation. Crombach used split-half reliability as his measure of deviation. Glaser used the number of items changed as his criterion. Lumsden used the proportion correct for a given item difficulty as his criterion.

Lumsden's work holds the most promise, but it has two serious drawbacks. It is based on population-dependent item difficulties and it appears to be only applicable to inter-person comparisons. However, it gives a firm idea for the solution of our problem. Lumsden uses the average person characteristic curve to define the reasonable, expected if you will, proportion of correct responses for the range of item difficulties on the test. By comparing the individual's person

characteristic curve, or proportion of correct responses, to the group's he is able to quantify the person's reliability as a function of the degree of departure from expectancy on an item by item basis.

In our original intuitive example this depature from expectancy was our basic reason for questioning the person fit of several individuals. We had an expectation of what a reasonable response vector should look like and we compared the real response vector to our imagined standard. What we needed was a quantitative method that could model the relationship between a person's ability and the difficulty of an item into the probability of a correct response. With this expected response for each item we could make the comparison between observed and expected responses in a systematic item-by-item fashion, combining these differences into a statistic to measure the person fit.

This approach seems to be straightforward and simple given our ability to create the expected values. Fortunately, we have that capability if we utilize the Rasch model and are able to formulate a variety of means of applying this basic idea to individual response strings.

III. THE MATHEMATICS OF PERSON FIT ANALYSIS

As we stated at the conclusion of the section II, the use of the psychometric model proposed by Rasch (1980) allows us to develop a comparison that is not possible in classical test theory, namely the comparison of expected and observed scores. The expected score under this model is a probabilistic expression based on two parameters, the difficulty of an item and the ability of the person. One form of Rasch's logistic probability function is shown in equation 1. In this equation β_v is the ability of a given person and δ_i is the difficulty of a given item.

$$(1)\quad P\{X_{vi}=1 \mid \beta_v, \delta_i\} = \exp(\beta_v - \delta_i)/(1+\exp(\beta_v - \delta_i)) = E(X_{vi})$$

where X_{vi} = 1 if the item is correct and 0 if the item is incorrect.

Since the parameters in this model are additive, it is possible to condition the effect of either parameter out of the estimation procedure. The estimators of both β_v and δ_i, which are derived by maximizing a conditional likelihood, are unbiased, consistent, efficient, and sufficient (Wright and Stone, 1979). The numerical properties of the conditional likelihood estimation, however, make the application of this

procedure difficult and costly for tests longer then 40 items. Wright and Panchapakesan (1969) developed a simple approximation for the conditional likelihood estimation. This unconditional likelihood estimation is computationally more efficient, particularly for longer tests, and provides a good approximation of the conditional method.

Using this "unconditional" method we are able to arrive at an estimate of each item's difficulty, σ_i , and an ability estimate, β_v , for each possible raw score, i.e., the total number of items answered correctly, on the test, except for zero and perfect raw scores, in a single calibration. The exclusion of the zero and perfect scores has a strong philosophical appeal given the desire to define a continuous variable which is only partially defined by the collection of items on a given test, but this property will have consequences in later applications.

In the Rasch model the raw scores are sufficient statistics for estimating the two parameters, person ability and item difficulty. From any raw score matrix we can derive an ability estimate for every person, provided that person got at least one item correct and one item wrong, as well as an estimate of any item's difficulty, provided at least one person got the item correct and at least one person got the item wrong. These estimates of item difficulty and person ability, expressed in the logit metric, enable us to utilize the

probabilistic function (1) which describes the interaction between the person and the item to calculate the expected response for any item-person interaction.

The probability defined by this function, in fact, represents the expected score of person v on item i . It has a value which ranges between 0 and 1. This function has a value of .5 when $B_v = \delta_i$ and approaches 0 and 1 when $|B_v - \delta_i| > 4.0$. Each item-person interaction has an observed value of 0 or 1 in the case of the dichotomously scored responses normally associated with achievement test data. The difference between the observed value and the expected value of each item-person interaction forms a residual which is the basis of the person fit analysis herein discussed.

The ability to summarize these residuals by a variety of fit statistics whose distribution can be approximately established is central to the previously expressed desire to be able to quantify the unexpected nature of unusual response patterns. There have been several attempts to formulate these fit statistics. One approach is the chi-square method. A chi-square with 1 degree of freedom is defined by (2):

$$(2) \qquad \chi^2_{(1)} = \frac{(Y-\mu)^2}{\sigma^2}$$

when Y is normally distributed with a mean of μ and a standard deviation of σ .

By definition the expected value of $\chi^2_{(1)}$ is one. In the

Rasch model, for items of difficulty σ_i and persons of ability β_v, the probability of a correct response is defined by equation 3.

$$(3) \qquad \Pi = \frac{e^{\beta_v - \sigma_i}}{1 + e^{\beta_v - \sigma_i}}$$

In this model X, the observed response, has an expected value equal to Π and a variance equal to $\Pi(1 - \Pi)$

$$(4) \qquad E(X) = \Pi \quad \text{and} \quad \sigma^2 = \Pi(1 - \Pi).$$

Using the concept of a standardized score where $\Pi(1 - \Pi)$ is written as W, then:

$$(5) \qquad Z^2 = \frac{(X - \Pi)^2}{W} = \chi^2_{(1)}$$

Z^2 is approximatly a chi-square with 1 degree of freedom.

From this we can create a series of fit statistics which are useful in the analysis of an individual's response string by summing over the almost independent chi-squares in a variety of ways. The first of these is a total fit statistic where Π is estimated by P which is based on the raw score derived from the total test and σ_i as estimated by d_i. Summing these Z^2 over all the items on a test we can form a chi-square with n-1 degrees of freedom as shown in equation 6.

$$(6) \qquad \chi^2_{(n-1)} = \sum_1^n Z^2 = \sum_1^n \frac{(X - P)^2}{W}$$

where n is the number of items.

This chi-square for a person can be converted into a mean square by dividing by the appropriate degrees of freedom, therefore

$$(7) \qquad V_1 = \sum_1^n \frac{(X_i - P_i)^2}{W_i} \Big/ (n-1) \qquad \text{where } E(V_i)=1$$

A cube root transformation can then be used to convert this mean square to a statistic which has a distribution very similar to the t-statistic (Wilson and Hilfery, 1931).

$$(8) \qquad t = (V_1)^{1/3}\left(\frac{3.0}{S_{V_1}}\right) + \left(\frac{S_{V_1}}{3.0}\right) - \left(\frac{3.0}{S_{V_1}}\right)$$

$$\text{where} \qquad S_{V_1} = \left(\Sigma\, 1/w - 4n\right)^{1/2} \Big/ n$$

The fact that this total fit statistic is very sensitive to unexpected (improbable) results led Wright (personal communication) to consider a more robust formulation of the total fit statistic. This robust statistic is based on the F statistic. Since for a binary response variable $E(X) = P$ and $Var(X) = P(1-P) \equiv W$. This $Var(X)$ is also equal to the expected value of the squared residual since $E(X-P) = 0$.

$$(9) \quad Var(X) = E(X-P)^2$$

We can then form the ratio of the two independent estimates of the variance of X , Var(X), which are accumulate over all of the items that a person attempts. This then is a F ratio with 1 and ∞ degrees of freedom.

This statistic is very similar in form to the chi-squared statistic in equation 6. The only difference is that in this statistic the numerator and the denominator are summed separately. The effect of this is to produce a statistic which is less sensitive to remote unexpected (improbable) responses. This robust statistic is often referred to as the weighted fit statistic while the chi-square in equation 6 is referred to as an unweighted fit statistic.

It is also possible to transform the weighted total fit statistic into a t-like statistic using the cube root transformation found in equation 8. In this case, however, S will take the form expressed in equation 10.

$$(10) \quad S_{V_t} = \left(\Sigma w - 4\Sigma w^2\right)^{1/2} \Big/ \Sigma w$$

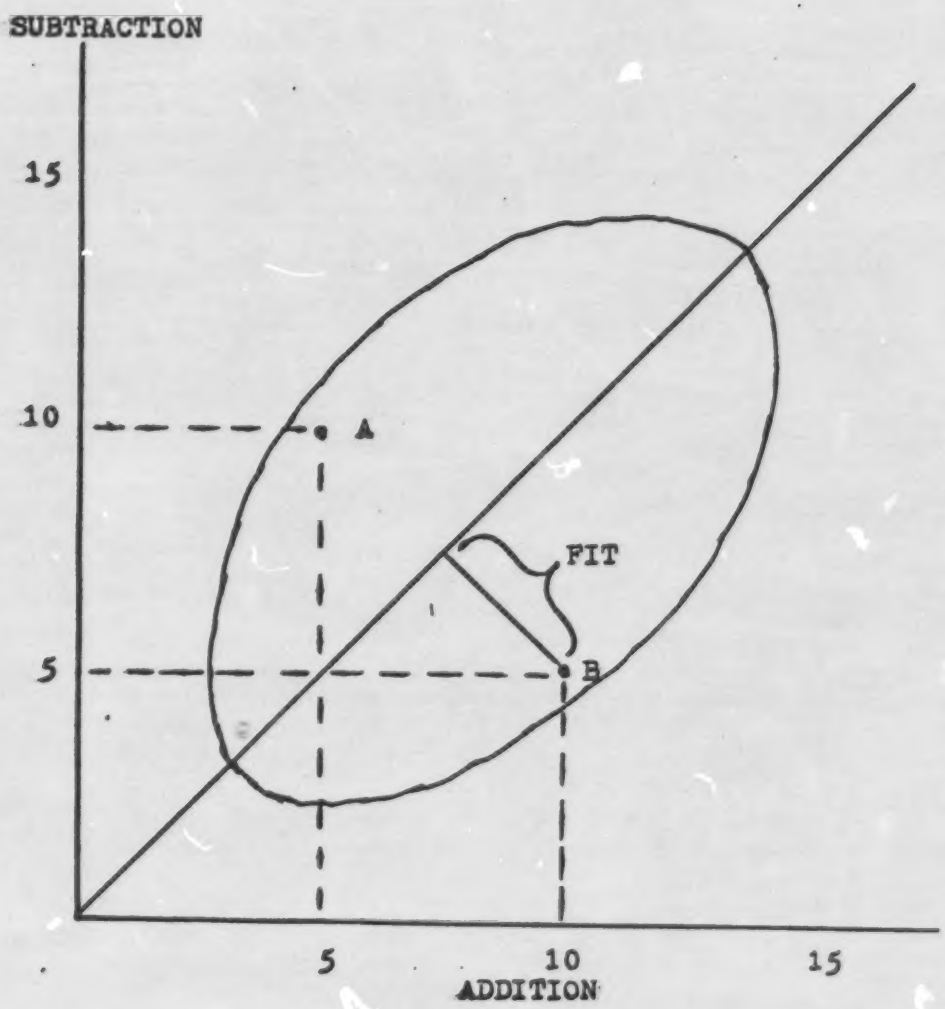
The computer programs, PANAL (Mead, 1978), used to produce the output discussed in detail in the subsequent sections of this paper are based on the more robust weighted total fit statistic discussed above and transformed to t-like statistic with an

expected value of 1 using the cube root transformation in equation 8 with the variance expressed in equation 10.

If we are to be able to completely understand the fit of a response string to the model we need to define two other types of person fit statistics. It is often the case that the items on a test can be divided into various meaningful subgroups based on item difficulty, item content, or position of the item on the test, to name just a few. Once we have specified these sets of items we can investigate the fit of the person within any set of items and the fit of the person between any related sets of items. The formulation of these two types of fit in terms of the weighted residual $\sum(X-P)^2 / \sum W$ is discussed below.

In the simple two subset case of a mathematics test made up of 15 addition and 15 subtraction items, the between set fit is, in fact, a quantification of the perpendicular distance a given individual lies from the identity line shown in Figure 1. Points lying far from the identity line in this figure indicate individuals for whom a sizable difference between raw scores on the addition and subtraction items exist and, assuming the item subsets are of equal difficulty, between the ability estimates based on each set separately. The critical notion is at what point does the total score based on all 30 items become an inappropriate representation of both the individual's addition and subtraction abilities. How far from the identity line does

FIGURE 1



an individual have to be before we would be reluctant to simply add up the raw scores from these two sets of items to estimate his ability?

The between-set fit statistic utilized in the PANAL program is an attempt at quantifying this idea. The between-set mean square used in this program is shown in equation 11.

$$(11) \qquad V_B = \frac{\sum_j \left(\sum_{i \in j} X_{ij} - \sum_{i \in j} P_{ij} \right)^2}{\sum_j \sum_{i \in j} W_{ij}}$$

where j= no. of sets of items and
i= items within sets

In this formulation the probability is based on the total-test ability estimate. This form of person fit analysis compares the observed score on any set of items with the score predicted by the individual's total-test ability estimate and the difficulty of the items that make up the set. This difference is then summed over the different sets that make up that group. If the person fits the model we would expect that these differences would be small.

The between-set fit statistic is primarily looking at the difference between the sum of the observed scores on a set of items, $\sum_i X_{ij}$ and the sum of the corresponding predicted scores, $\sum_i P_{ij}$. This difference is squared, summed over the number of groups and divided by the sum of the expected values (variance), $\sum_j \sum_{i \in j} W_{ij}$, of the values in the numerator. This

produces a mean square which again may be transformed using the cube root transformation in equation 8 which has the variance found in equation 10.

The within-set fit statistic utilized in the PANAL program is an attempt to quantify the unexpected (improbable) responses within a given subset of items. The formulation of the within-set fit statistic is found in equation 12.

$$(12) \qquad V_w = \frac{\sum_i (X_i - P_i)^2}{\sum_i W_i}$$

where i= items within a given set.

Once again the within-set fit statistic is based on the total test ability estimate and not on an ability estimate derived only from the items in that subset.

The primary reason for this is the inability to estimate abilities for zero and perfect scores. Zero and perfect total test scores are relatively infrequent in comparison to the frequency of zero and perfect scores on the smaller subsets of items. Since the ability of the individual is necessary to calculate the fit statistic, the greater frequency of zero and perfect scores on the subsets is problematic. In the case of the total test these zero and perfect scores are not as problematic since by default both of these cases fit the model. The presence of zero or perfect subset scores will not effect

the calculation of the fit statistic based on the total test. But it would have made it impossible to calculate the within fit statistic if the subset ability estimates had been used to calculate the within-set fit.

The consequence of using the total test abilities to calculate the within-set fit statistic is that the within-set fit statistic is sensitive to possible variation of the abilities as measured by the individual subsets as well as the within-set variation it is primarily designed to measure. In this formulation the between-set and within-set fit cannot be thought of as an exact partitioning of the total test fit in the classic analysis of variance sense . This fact can cause some problems in the interpretation of the within-set fit statistic.

Consider the case of two individuals with an ability of 2.0 logits. This ability is based on a test which contains a subset of five items whose difficulties are as follows: -2.0, -1.5, -1.0, -0.5, 0.0. Both of these individuals missed one of the five items. The first individual missed the easiest item, difficulty -2.0, and had a response string for this subset that looked like 01111 . The second individual missed the hardest item in the subset, difficulty 0.0, and had a response string for this subset that looked like 11110. These two cases represent the best and worst possible cases of fit for an individual with a score of 4 on this subset of items. Using the

within-set fit statistic in equation 12 on these five items and transforming this mean square using the cube root transformation we find that the worst fitting individual has a within-set fit of 1.48 and the best fitting individual had a within-set fit of 1.30. The majority of the misfit in both cases is due to the fact that the one item missed was at least 2 logits below the two individual's total test ability and thus this fit statistic does not accurately reflect the response patterns within the subset of items. We can see that the within-set fit statistic which we will be using really is not a true within-set fit statistic. It is sometimes influenced by the between-set variation introduced by using the total test ability estimate rather than the within-set ability estimate. This will have only a small impact on our analysis of the person fit data, but it should be kept in mind as we consider the individual cases.

IV. EXAMPLES OF PERSON ANALYSIS

So far we have discussed the intuitive concept of the quality of a measure for a given individual and reviewed the attempts at quantifying and detecting this quality over the last 60 years. We, also, have discussed the applicability of the Rasch model to the quantification of the quality of a measure, as well as the various ways of utilizing the properties of the Rasch model to detect occurrences of low quality measurement. We have discussed the mathematics utilized in the computer program, PANAL (Mead, 1978), which is used to analyze the individual response strings in this study for the presence of low measurement quality or misfit. However, we have yet to demonstrate either the practical applicability of this concept or the degree to which the occurrence of misfit affects the ability estimate for an individual. Finally we have yet to discuss the potential remedies that are available to extract from a misfitting response string a valid ability estimate.

The best way to do this is to look at some real people, twelve in all, analyze their response vectors to determine why the fit statistics characterize their responses as misfitting and see if there is anything that can be done about extracting a valid ability estimate from these response vectors. There

are five general categories of misfits which we will attempt to detect; guessing, startup (fumbling), plodding, sloppiness, and problems involving item-person interactions.

Guessing is the result of an individual entering random responses to items. Startup (fumbling) refers to a tendency among certain individuals to do poorly on the initial items on a test, but then, as the test progresses, to settle down and perform at a higher level of ability. Plodding refers to a test taking strategy exhibited by certain age groups and cultures. It involves working very deliberately and making certain that each item is as correct as possible before proceeding on the next item. As a result these individuals often do not finish a timed test, leaving the later items blank. Sloppiness (carelessness) refers to the tendency of some individuals to carelessly mark responses on answer sheets. Detection of this type of response pattern is often limited only to able individuals who incorrectly mark easy items. This type of response pattern may also result from the tendency of some able individuals to over-intellectualize a very easy item so as to overlook the obvious for a more subtle distinction. Item-person interactions refer to the interaction between an individual and some definable group of items. These items may be similar in terms of their content, format, or bias regarding age, sex, or cultural background.

The individual cases used to illustrate these types of

misfit were drawn from the 46,000 freshmen entering New Jersey public colleges in the fall of 1978. Each freshman was required by the Department of Higher Education to take a placement test prior to enrolling in the college. The test battery, the New Jersey College Basic Skills Placement Test, consists of five objective sections and an essay. The five objective sections are Reading Comprehension, Sentence Structure, Logical Relationships, Computation and Elementary Algebra. The examples used here were specifically chosen from the Logical Relationships and Elementary Algebra tests. These two represent the easiest and hardest tests, respectively, in the battery. The average state-wide score on the Logical Relationships test is 78% correct while the average state-wide score on the Elementary Algebra test is 47% correct. The Logical Relationships test contains 50 four choice multiple choice items with four content subscores, categorizing ideas, connectives, analogies, and principles of organization. The Elementary Algebra test contains 30 four choice multiple choice items with two content subscores, solution of equations and operations with algebraic expressions.

The comparison of these content subscores will constitute one of the three between-subset fit statistics calculated for these individuals. The second between-subset fit statistic, sequence, will be based on a comparison of the first and second half of the Logical Relationships test and the 3 thirds of the

Elementary Algebra Test. The third between-subset fit statistic, difficulty, will be based on a grouping by item difficulty into four groups high, high middle, low middle, and low. We will discuss various other aspects of these two tests as we review the individual cases.

IV A. GUESSING

Person #1 had 22 correct responses to the 50 items on the Logical Relationships test. This yielded an ability estimate of -0.29 logits. The ability estimate (measure), its standard error, the total fit statistic and the three between subset fit statistics are found in the upper left hand corner of page A-1. This individual was characterized as a potential misfit due to the high total fit statistic (1.8) and the high between set fit statistic (1.6) for difficulty. Experience with these tests indicates that individuals with fit statistics larger than 1.2 to 1.5 are worthy of attention. The choice of these values is motivated by a desire to detect as much of the real misfit as possible, even if it means having to review cases where the fit statistics fall into this range simply by chance. Individuals with fit statistics that are large by chance can always be reported as having acceptable results, if a review indicates that there is no systematic misfit.

To the right of the ability estimate and the fit statistics is a plot of the mean residuals (X - P) from each item, or group of items by difficulty. The solid vertical line in the center represents the individual's ability (-0.29) while the two vertical lines of dots represents the individual's ability ±1 logit. The item difficulties in logits are listed across the bottom of the graph. In this graph the distance below the horizontal dash line, labeled 0.0, represents our

surprise, the unexpected nature, at an incorrect response. The distance a point lies above the horizontal center line represents our surprise at a correct reponse. The further the point is from the line the more unexpected the response. The number plotted indicates the number of items that were grouped at that particular item difficulty. In these cases the point plotted represents the mean surprise for all the items at this difficulty.

If we examine the points on this plot we can begin to see the signs of some unexpected test taking behavior. This individual missed the second easiest item, difficulty -1.8, while getting the second hardest item, difficulty 1.8, correct. The probability of getting the easy item incorrect for a person of this ability, -0.29, is 0.20. The probability of this person getting the hard item correct is 0.11. The total fit statistic, which summarizes the surprise plotted in this graph, correctly calls attention to this person. If we had simply been reviewing the dichotomous responses to these items listed in difficulty order we would have probably selected this person for further attention. The task is much easier using the fit statistics. Instead of having to look at five different response vectors, one for each of the five tests in this battery, for 46,000 people (230,000 response vectors in all) we can have the computer review them and print out those that we would like to see.

The mean residual plot and the total fit statistic do not tell the whole story though. There is more of interest to be found by looking further at this individual. Directly below the mean residual plot on page A-1 we have the statistics and some plots for the three subsets which we have defined for this particular test. The first one listed is subscores, our code word for the four content sets on the Logical Relationships test. Items 1-12 deal with categorizing ideas. Items 13-25 deal with connectives. Items 26-37 deal with analogies. Items 38-50 deal with problems of organization. We have labeled these sets sub1, sub2, sub3, and sub4, respectively. The number of items in each set is listed in the second column, labeled count. The individual's raw score on each set is listed in column 3 labeled score. The fourth column lists the individual's ability in logits. The first entry in that column, -0.29, is the individual's total ability. Directly below that are the abilities for the four sets. These abilities are based solely on the items in that set. The standard error of measure for the total ability and the four set abilities are listed in the fifth column, labeled S.E. The sixth column, labeled WITHIN, lists the within set fit statistic for the entire test which is in fact the total test fit statistic. The last column, labeled BETWN, lists the between set fit statistic. The between fit statistic is a measure of the misfit due to groups (sets) of items. The

figures listed in the top row of this table are repeats of the statistics listed at the top of the page.

The plot to the right of the subscores table is a graphical display of the total ability estimate, dotted vertical line, and the four subscore set ability estimates, asterisks. One standard error in either direction is indicated by the horizontal dashes. This plot summarizes the information in columns 4 and 5 in a visual form.

The plots for the other two subset configurations contain the same type of information, only this time it is based on a different grouping of the items. Sequence compares the first half of the test with the last half. Difficulty compares sets of items which are grouped by item difficulty: easy (diff. < −0.75), low middle (−0.75 < diff. < 0.0), high middle (0.0 < diff. < 0.8), and hard (diff. > 0.8). The selection and use of subsets is a very important part of analyzing individual response vectors. Each different configuration is sensitive to special types of misfit. Difficulty is useful in detecting guessing and sloppiness. Sequence often picks out startup and plodding. Subscores (content) is sensitive to item content-person interactions.

Let us turn our attention to the information that these three set analyses provide us about person 1's problem. Looking at subscores we can see that one of the ability estimates, sub3, is at least 1.3 logits lower than the other

three ability estimates in that subset. Despite the fairly large difference between these ability estimates, the between set fit (0.6) is still below the critical level that we set. This highlights one of the reasons for choosing low critical values for fit statistics. The between set fit statistic appears to be rather robust and will allow fairly large between set differences in ability when the fit statistic is less than one. If we look at the within set fit statistic for each of the four sets in subscores we will note that sub4 has a within set fit of 1.4. This suggests that the responses to the items in sub4 are not consistent with what the Rasch model would predict for this person given the total ability for the person.

We should note that the within set fit statistic is calculated by using the total ability estimate rather than the set ability estimate. This is done to avoid the problem with zero and perfect scores. The Rasch model cannot estimate ability for individuals with zero or perfect scores. In this analysis, it is easy to screen individuals for zero or perfect total test scores. In most cases this screening process will not eliminate many individuals from the analysis. However, when you begin looking at ten 12 to 25 item subsets, as we are in this example, the chances of finding zero or perfect subset scores is increased. By choosing to use the total ability to calculate the within set fit statistics, we will be able to estimate the individual's within and between set fit

statistics. We can also estimate an individual's ability on any set in which he does not have a zero or perfect score. Since these statistics are the most important point of this type of analysis, this would appear to be a reasonable procedure.

The low ability estimate for sub3 and the high level of inconsistency of responses within sub4 (within fit = 1.4) begin to give us the impression that something different is going on in the last 25 items than was going on in the first 25 items. If we look at the sequence sets we again see that the individual's ability is 0.85 logits lower on the second half than on the first. The within set fit for the second half also indicates an unacceptable level of inconsistency for these responses, too many easy items wrong and too many hard items right. Again this information reinforces our notion that something is different about the second half of this test. If we look at the difficulty subsets we can see that the bewteen set fit statistic (1.6) is above the critical level that we have set. This person's ability estimate based on the 9 easiest items, -1.5, is 2.1 logits below the ability estimate based on the 15 high middle difficulty items, 0.6. An inspection of the number correct (score) for these two sets confirms this. Why should a person score 4 out of 9 on the easiest items and 8 out of 15 on much harder items? The difference in these two ability estimates begins to make our

total ability estimate look unrealistic. What is the individual's true ability: -1.5, -0.29, or 0.6?

Notice, also, that even though individual 1 only answered 2 of the 8 hard items correctly, there was an inconsistency within that set, fit=1.6. Where does all of this leave us? We know that there is an inconsistency in the total test response vector, but if we divide the test in half, the first half is consistent and the second is not. We know that the percentage correct drops from 56% on the first half to 32% on the second half and that this individual missed an unexpected number of easy items. Can we identify this individual's problem?

If we look at page A-2 the problem should become clear. Page A-2 is a plot of the log logit residuals for each item in sequence order. The A plotted for the first 12 items indicates membership in sub1, B indicates membership in sub2, C membership in sub3, etc. Points plotted on or above the center line, labeled 0.0, indicate that the item was correct. Points below the line indicate that the item was incorrect. The distance above or below the line indicates the amount of surprise, the unexpectedness, of that response. If we look at this plot, some of our questions should be answered. The two items which originally drew our attention to this person, the second hardest and the second easiest, are items 31, lowest C, and item 42, highest D. They both occur on the second half of the test, the half which we felt was unusual.

Can we identify the problem? We must remember that any attempt to do so is speculative but the pattern here is fairly typical. It would appear that individual 1 seriously attempted items 1-25. For whatever reason individual 1 appears to have guessed at items 26-50. Maybe this individual ran out of time and just tried to fill up all of the spaces at the end of the test, maybe the individual became frustrated and quit working, maybe this individual copied the answers to the first 25 items and couldn't copy the last 25 and so he guessed. Whatever the reason, the analysis of the responses indicates a serious attempt at the items on the first half but not on the second half. The steplike pattern from low ability on the easy item set to high ability on the hard item set seen in the graphical display of the difficulty subset analysis is typical of guessing. An alternative hypothesis might suggest that individual 1 was not guessing on the second half, only that this individual didn't know the content of these two content subsets, i.e. we had a person/content interaction. If individual 1 had had a better set fit on difficulty, this might have been a possibility , but the inconsistency of the responses in the second half argues against that theory.

Can we do anything about this individual's ability estimate? Clearly there are two alternatives. We can under no circumstances use the last 25 items on the test to estimate individual 1's ability. We can accept the first 25 items as an

accurate description of the person's ability or we can retest the person. This decision process might be aided if we were able to perform this analysis close to the time of the testing session and talk with this person about the test. Unfortunately we cannot, so we must do the best with what we have. My inclination would be to accept the first half of the test as an adequate representation of this individual's ability. This individual's ability based on the first half would be 0.14 logits as opposed to -0.29 based on the whole test. The standard error for the ability based on the first half is only slightly higher than the total test standard error, 0.44 vs. 0.31.

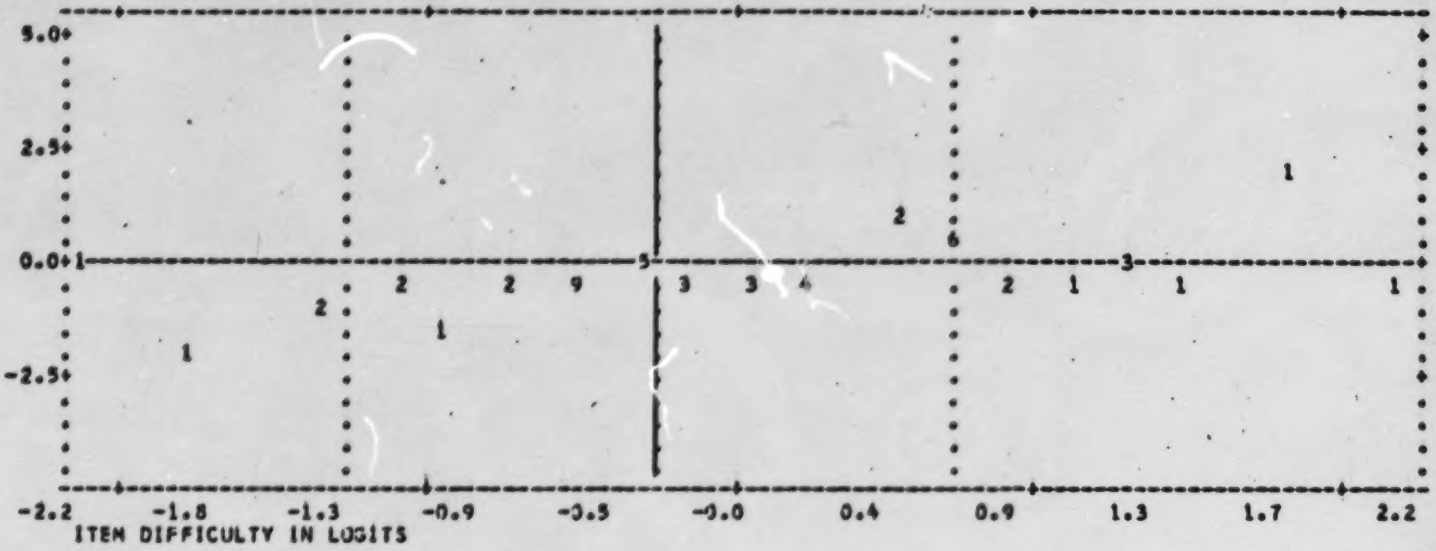
PANAL 3 -- SELECT T GREATER THAN 1.5

PERSON IDENTIFICATION 1 PAGE A-1

SUMMARY STATISTICS

MEASURE	-0.29
SEM	0.31
TOTAL T	1.8
BETWEEN SUBTEST SUBSCORES	0.6
SEQUENCE	0.9
DIFFICULTY	1.6

MEAN RESIDUAL BY DIFFICULTY



NAME	COUNT	SCORE	ABILITY	S.E.	WTHIN	BETWN
SUBSCORES			-0.29	0.31	1.8	0.6
SU81	12.	8.	-0.06	0.64	0.9	
SU82	13.	6.	0.30	0.60	0.8	
SU83	12.	3.	-1.41	0.70	0.9	
SU84	13.	5.	-0.09	0.60	1.4	

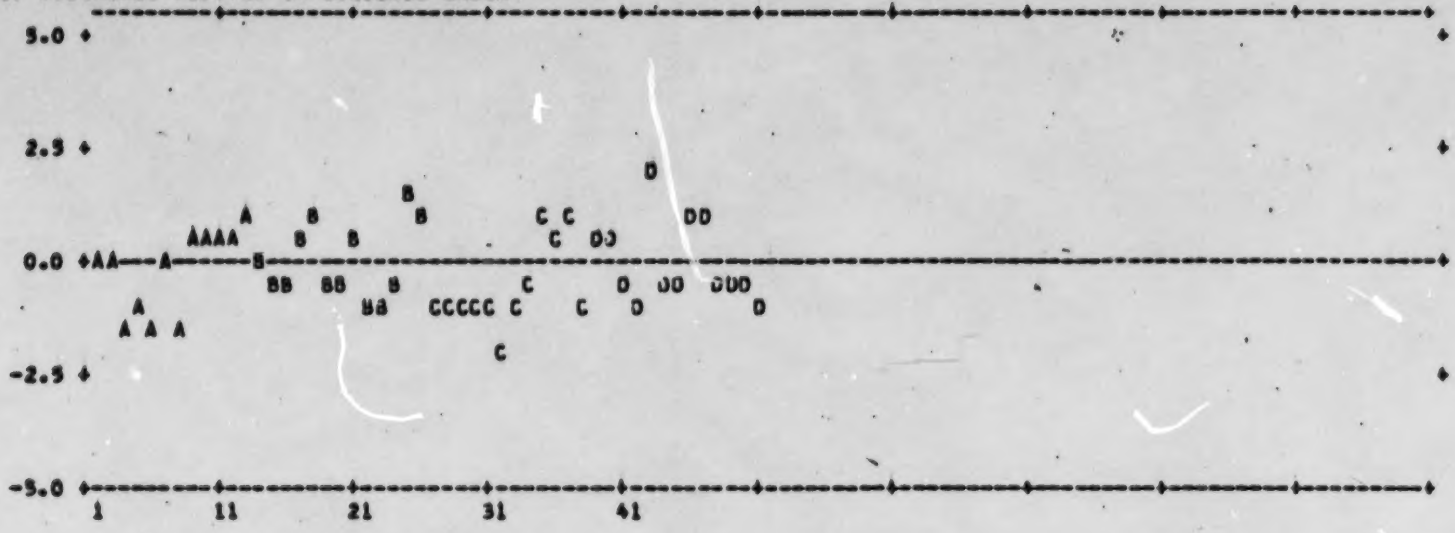
NAME	COUNT	SCORE	ABILITY	S.E.	WTHIN	BETWN
SEQUENCE			-0.29	0.31	1.8	0.9
1HLF	25.	14.	0.14	0.44	1.0	
2HLF	25.	8.	-0.71	0.45	1.7	

NAME	COUNT	SCORE	ABILITY	S.E.	WTHIN	BETWN
DIFFICULTY			-0.29	0.31	1.8	1.6
-.75	9.	4.	-1.50	0.69	0.9	
0.00	18.	8.	-0.60	0.48	0.2	
0.80	15.	8.	0.60	0.52	0.9	
2.00	8.	2.	0.28	0.81	1.6	

PANAL 3 — SELECT T GREATER THAN 1.5

PERSON IDENTIFICATION 1

LOG LOGIT RESIDUALS PLOTTED IN SEQUENCE ORDER



Individual 2 answered 8 of the 30 items on the Elementary Algebra test correctly. This yielded an ability estimate of -1.16 logits. Looking at the fit statistics on page A-3 we see individual 2 has a total fit statistic of 2.0, a sequence between-set fit of 2.1, and a difficulty between-set fit of 7.7. The number of correct items for individual 2 might lead us to immediately suspect guessing but we should remember that this is a very difficult test and nearly 40% of the people taking this test scored in the chance range, 10 or less items correct. Looking at the mean residual by difficulty plot, top right page A-3, we can see that individual 2 missed all six of the easiest items on the test yet got the second hardest right. In all, this individual answered six items correctly that were more than 1 logit above his ability, a rather unusual occurrence.

Looking at the subscore set fit statistics, we see that even though there is little between-set difference, fit=-2.7, there is considerable inconsistency in the within-set responses to sub1, fit=2.3. For the sequence set analysis we find a between set fit of 2.1. This is a chance result which sometimes occurs in cases of guessing. Again two of the three within-set fit statistics have large values, 1.9 and 1.3. In the case of the difficulty set analysis we see a between-set fit statistic of 7.7. This is the result of missing all of the easy items and correctly responding, guessing?, to 5 of the 11

high middle difficulty items. If we look at the four within-set fit statistics for difficulty we see that three of the four values are large. In all, 9 of the 13 fit statistics reported in this analysis have a value greater than 1.1. It would be difficult to find a more obvious case of guessing.

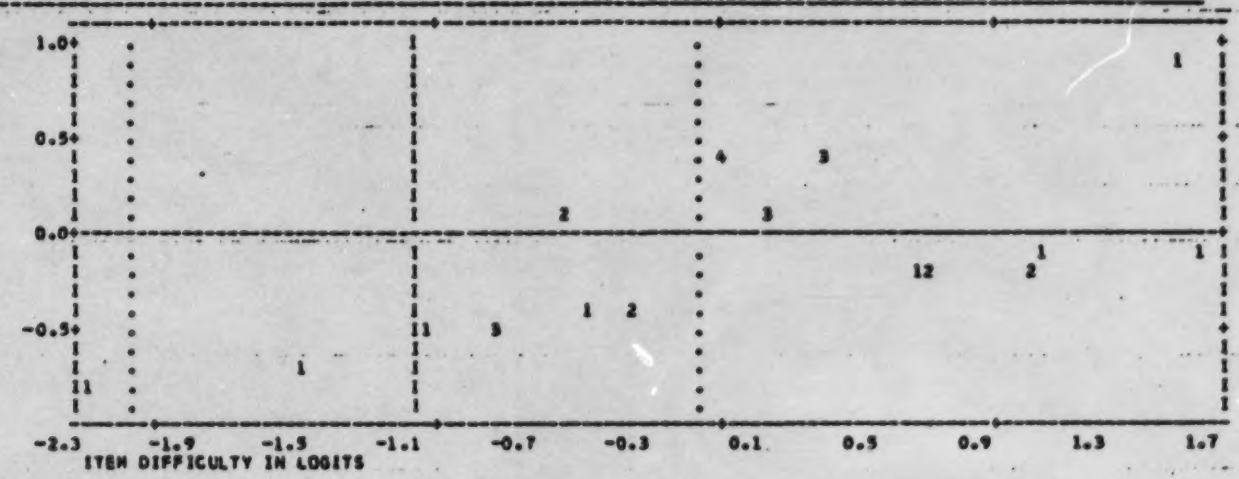
This pattern is typical of guessing. There is obviously no part of this response vector that can be considered valid. There is little else we can do in this instance other than to require that this individual be retested, hopefully after a little counseling.

NJCSSPT ELEMENTARY ALGEBRA TOTAL T GREATER THAN 1.5

PERSON IDENTIFICATION 2 PAGE A-3

SUMMARY STATISTICS	
MEASURE	-1.16
SEM	0.44
TOTAL T	2.0
SKIPPING OMITS	
MEASURE	-1.10
SEM	0.45
BETWEEN SUBTEST	
SUBSCORES	-2.7
SEQUENCE	2.1
DIFFICULTY	7.7

MEAN RESIDUAL BY DIFFICULTY



NAME	COUNT	SCORE	ABILITY	S.E.	WTHIN	BETWN
SUBSCORES			-1.16	0.44	2.0	-2.7
			-1.10	0.45		
SUB1	17.	6.	-0.75	0.54	2.3	
SUB2	12.	2.	-1.78	0.89	0.6	

NAME	COUNT	SCORE	ABILITY	S.E.	WTHIN	BETWN
SEQUENCE			-1.16	0.44	2.0	2.1
			-1.10	0.45		
1TEN	9.	2.	-1.87	0.88	1.9	
2TEN	10.	2.	-1.72	0.81	0.2	
3TEN	10.	4.	0.12	0.67	1.3	

NAME	COUNT	SCORE	ABILITY	S.E.	WTHIN	BETWN
DIFFICULTY			-1.16	0.44	2.0	7.7
			-1.10	0.45		
-.75	6.	3.	.**********		1.3	
0.00	7.	2.	-1.26	0.84	0.0	
0.80	11.	5.	0.14	0.61	1.7	
2.00	5.	1.	-0.10	1.08	1.2	

IV B. STARTUP (FUMBLING)

Person #3 had 41 correct responses to the 50 items on the Logical Relationships test. This yielded an ability estimate of 1.76. The total test fit statistic for this individual was 3.2. In looking at the mean residual by difficulty plot on page A-5 we can see that this person missed 3 of the 6 easiest items on the test. These items were at least 2.5 logits below this person's ability. The probability of missing an item that far below a person's ability is 0.08 or less. Other than these three items, the rest of the residuals are very close to the expected values, i.e. the points are plotted close to the 0.0 axis.

Looking at the subscore set analysis, we find a marginal value for the sequence between set fit, 1.1. The ability estimate for this subset indicate that person 3's ability on the first set is between 1.2 and 2.0 logits below the ability estimate based on the other three sets. The within set fit statistic for the first set, sub1, is very high, while the within set fit statistics for the other three sets are acceptable. Looking at the sequence set analysis, we see that the two ability estimates are considerably closer and that the between set fit is lower, -0.1. The high ability for sub2 has lessened the impact of the low ability estimated by sub1, producing a middle value which is not too different from the values of sub3 and sub4. As in any instance where two

considerably different ability estimates result from diffferent item groupings, it is questionable whether the average of the two different abilities is a meaningful indication of the individual's true ability.

If we look at the within set fit statistics for subscores and sequence we notice that the fit statistic for the set containing the first 12 items is high in both analyses: 7.3 in subscores and 6.4 in sequence. Even though combining sub1 and sub2 produced an average ability close to the second half ability, the inconsistency of the responses remained high. The difficulty set analysis helps us to further isolate the problem. It confirms the information we gathered by looking at the mean residual by difficulty plot. The between set fit statistic is high, 3.1. We can see that the ability estimate based on the easy items is 2.5 to 3.0 logits below the estimates for the other three difficulty sets. The within set fit statistic for the easy set is also high, 7.0.

The residual plot in sequence order gives us the final support for characterizing this person as having trouble starting the test. This person missed 3 of the first 7 items on the test. These also happen to be among the easiest items on the test. After this initial inconsistency, the individual settled down and performed very well on the rest of the test, missing an occasional question but otherwise doing well. One important question is how can we be sure that this is startup

and not the fact that this person was just less able with regards to the content of subi. This would be the situation that we have characterized as a content/person interaction. The ability estimates in the subscores and sequence analyses might point to a content/person interaction if it were not for the fact that the within set fit statistics for the sets containing the first seven items were so high. If this individual did not know the content of subi as well as the content of the other three sets we would have expected this individual to answer the easy items correctly and miss the difficult items. This would have yielded an equivalent ability estimate but the within set fit statistics would have also been acceptable. It is also unlikely that the within set fit for subi was the result of guessing, since the number correct is substantially above the chance level. This type of residual pattern occurs fairly frequently among higher ability individuals.

The question remains as to what is this individual's correct ability. The 1.76 seems to be an underestimate resulting from a slow or erratic start on the test and most likely is not a true reflection on this individual's ability. Retesting this individual simply to correct for three items seems extreme. The best alternative would seem to be to simply omit these three items from the test and base this individual's ability on the remaining 47 items. This would yield an ability

estimate of approximately 2.31.

PANEL 3 -- SELECT BS GREATER THAN 1.5

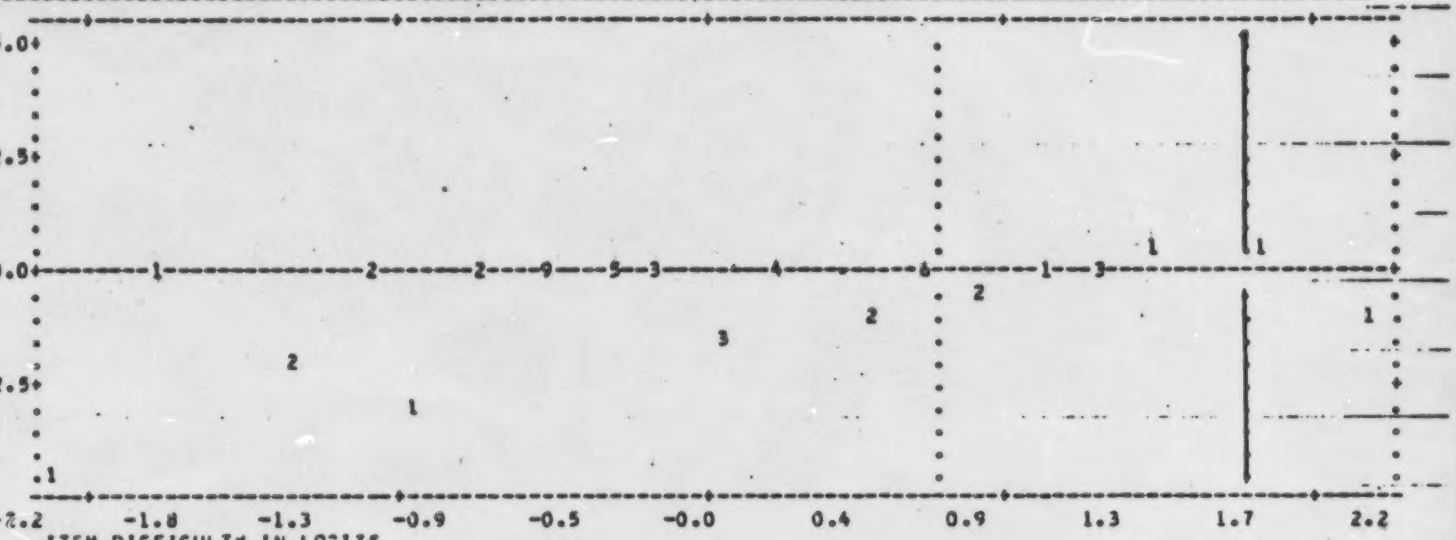
PERSON IDENTIFICATION 3 PAGE A-5

SUMMARY STATISTICS

MEASURE 1.76
SEM 0.39
TOTAL T 5.2

BETWEEN SUBTEST
SUBSCORES 1.1
SEQUENCE -0.1
DIFFICULTY 3.1

MEAN RESIDUAL BY DIFFICULTY



NAME	COUNT	SCORE	ABILITY	S.E.	WITHIN	BETWN
SUBSCORES			1.76	0.39	5.2	1.1
SUB1	12.	9.	0.39	0.69	7.3	
SUB2	13.	11.	2.41	0.80	-1.9	
SUB3	12.	10.	1.62	0.81	0.6	
SUB4	13.	11.	2.28	0.79	-0.6	

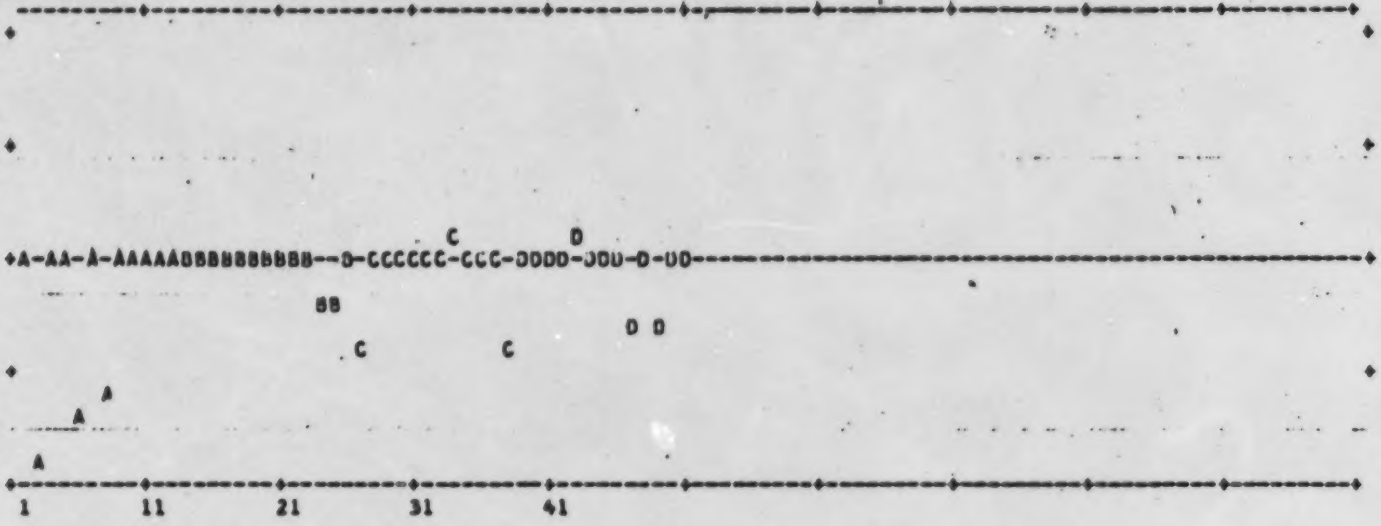
NAME	COUNT	SCORE	ABILITY	S.E.	WITHIN	BETWN
SEQUENCE			1.76	0.39	5.2	-0.1
1HLF	25.	20.	1.51	0.54	6.4	
2HLF	25.	21.	1.99	0.57	-0.1	

NAME	COUNT	SCORE	ABILITY	S.E.	WITHIN	BETWN
DIFFICULTY			1.76	0.39	5.2	3.1
-.75	9.	6.	-0.54	0.72	7.0	
0.00	18.	17.	2.47	1.03	-2.0	
0.80	15.	13.	2.36	0.76	-0.3	
2.01	8.	5.	1.94	0.74	0.4	

PANAL 3 -- SELFCT 83 GREATER THAN 1.5

PERSJN IDENTIFICATION 3

LJG LOGIT RESIDUALS PLUTTED IN SEQUENCE URDER



C2

C1

Individual 4 answered 20 of the 30 items on the Elementary Algebra test correctly. This yielded an ability estimate of 0.81. Looking at the fit statistics on page A-7 we see this individual has a high total test fit statistic, 3.3. Inspection of the mean residual by difficulty plot indicates this person missed the two easiest items on the test but answered 2 of the 5 most difficulty correctly. The sequence analysis between-set fit statistic is high, 4.2. Looking at the three sequence sets we see that the ability estimate for set 1 is 3.4 logits lower than the other two sets. The within set fit statistic is also high for the first set, 4.8. This would indicate a high degree of inconsistency within this set.

The difficulty set analysis indicates that this individual has a lower ability estimate for the easy set of items than for the other three sets. The within set fit statistic for the easy item set is also high, 5.6, indicating a high degree of inconsistency. Contrast this with the within set fit for the hard item group, 0.6. In both instances this individual had approximately the same percentage of items correct, but in the latter case responses match our expectations, i.e. the most difficult items were incorrect while the least difficult items were correct.

The residual plot in sequence order confirms our suspicion that this is a case of startup. In fact this individual missed 5 of the first 6 items on the test even though these were the

easiest on the test. If this is startup, it is an extreme case. It is hard to imagine an individual going from 3 of the first 10 items correct to 9 of the next 10 correct as the items get more difficult and with no major change in the content of the items. One explanation might be that after doing the first 6 items this individual realized that he was unable to do the items and copied from someone else. The ultimate outcome of this action would have placed him into a more difficult math course, so that explanation seems fairly implausible but not impossible.

What action should we take in this case? In the first case we recommended dropping the three problem items. In this case we would be dropping almost 1/3 of the items on the test and in doing so we would raise this individual's ability from 0.81 to 2.07, quite a jump. Perhaps the first step in resolving this dilemma would be to see if this individual did the same thing on the other four tests in the battery. It would also be useful to talk with this individual and find out what their perceptions of the test were; maybe there were external circumstances in the test administration which might explain the sudden change. If after talking with this individual I was convinced that the problem was really startup, I would use the 2.07 ability estimate and recommend a test taking skills workshop. Otherwise I would be inclined to rest this individual.

PANAL 3 -- SELECT T GREATER THAN 1.5

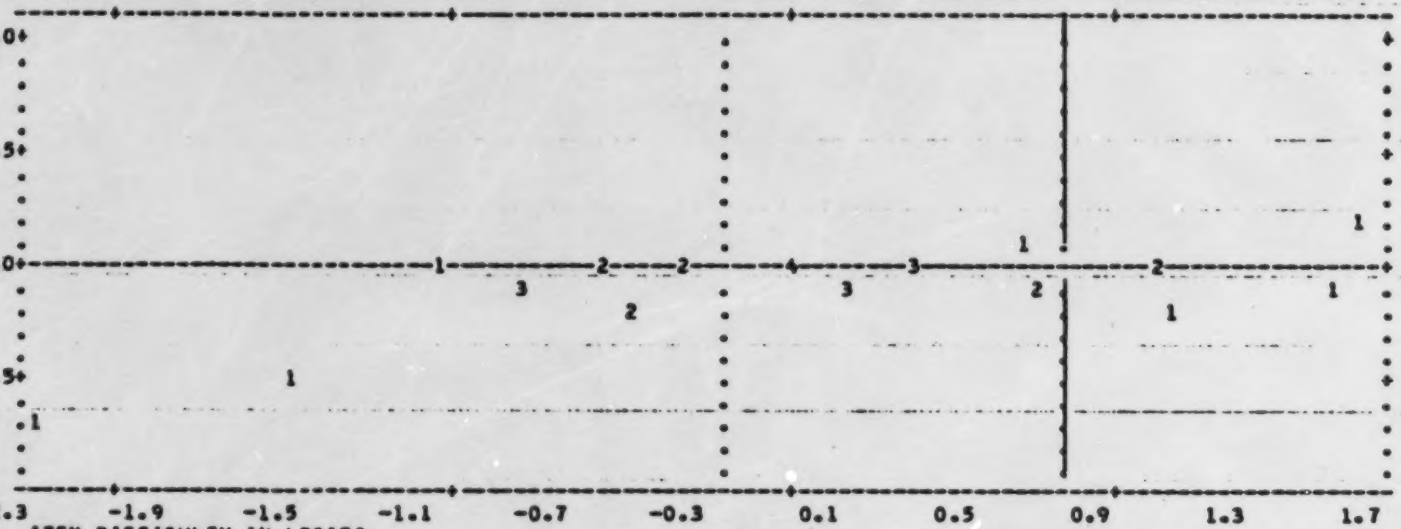
PERSON IDENTIFICATION 4

SUMMARY STATISTICS

MEASURE 0.81
SEM 0.41
TOTAL T 3.3

BETWEEN SUBTEST
SEQUENCE 4.2
SUBSCORES -9.6
DIFFICULTY 1.9

MEAN RESIDUAL BY DIFFICULTY



NAME	COUNT	SCORE	ABILITY	S.E.	MTHIN	BETWN
SEQUENCE			0.81	0.41	3.3	4.2
ITEM	10.	3.	-1.36	0.76	4.8	
2TEM	10.	9.	2.15	1.08	-4.7	
3TEM	10.	8.	2.05	0.81	-0.8	

NAME	COUNT	SCORE	ABILITY	S.E.	MTHIN	BETWN
SUBSCORES			0.81	0.41	3.3	-9.6
SUB1	18.	12.	0.74	0.54	1.8	
SUB2	12.	8.	0.91	0.64	3.3	

NAME	COUNT	SCORE	ABILITY	S.E.	MTHIN	BETWN
DIFFICULTY			0.81	0.41	3.3	1.9
-.75	6.	3.	-1.21	0.84	5.6	
0.00	8.	7.	1.61	1.07	-0.6	
0.80	11.	8.	1.33	0.68	-0.3	
2.00	5.	2.	0.89	0.92	0.6	

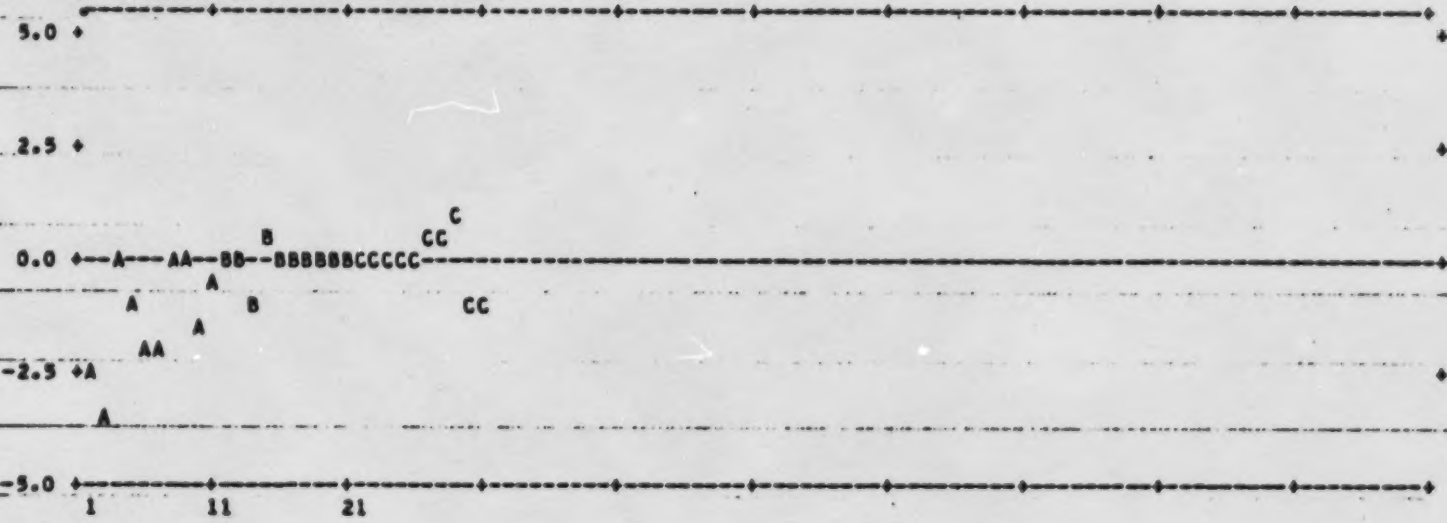
C5

C6

PANAL 3 -- SELECT T GREATER THAN 1.5

PERSON IDENTIFICATION 4

LOG LOGIT RESIDUALS PLOTTED IN SEQUENCE ORDER



IV C. PLODDING

Person #5 answered 41 of the 50 items on the Logical Relationships test correctly. The ability estimate for this individual was 1.76 logits. Looking at the total test fit statistic on page A-9 we see that this individual was a little more consistent than the probabilistic model would have us to expect. This individual missed only 1 of the 27 lower difficulty items on the test. We can see how closely the residuals are clustered about the center line in the mean residual by difficulty plot. There is nothing in this plot or in the total fit statistic which would lead one to expect any problems with this measure. This is the first case that we have looked at where the total test fit statistic was less than 1.5.

If we look at the subscores set analysis we can begin to detect a problem. The between set fit statistic is 2.4. Two of the four subsets have all correct responses, which means that we are unable to estimate the individual's ability on these two sets. The 9999.90 in the ability and standard error columns indicates our inability to estimate the set ability. The ability on a perfect set is higher than the highest possible ability on that set, so the ability is plotted on the extreme right hand side of the subscores ability plot. The ability estimate for sub4, 0.25, is at least 2.1 logits lower than the other three ability estimates. The within set fit statistic

for the fourth set is also high, 1.6. The sequence set analysis confirms this. The between set fit is 1,4 and there is a 1.5 logit difference in the two ability estimates. In some sense the sequence set analysis is redundant for the Logical Relationships test, but this is due to the unusual construction of the test. Since the items in each of the content subsets are grouped together, the subscores fit statistics are picking up many of the problems that sequence analysis would find if the four content types were mixed on the test. This is why startup and plodding are reflected in the subscores fit statistic on the Logical Relationships test but not on the Elementary Algebra test.

Finally, if we look at the difficulty subset analysis we find about a one logit difference between the ability estimates, but overall the fit statistics indicate that the differences are about what the model would predict. What then is the problem? Looking at the residual plot in sequence order on page A-10, we see a long string of correct responses, 41 correct in 43 items, followed by 7 incorrect items. An analysis of the original responses indicated that in fact these 7 items were not attempted. The last two items were relatively easy items for a person of this ability, so it is improbable that this individual could not answer them. It is most likely that this person ran out of time. This would account for the high within set fit statistic on sub4. This person answered

the early hard items correctly but was marked incorrect on the easy items that were not reached. This produced the inconsistent pattern for sub4.

This type of response pattern is best characterized as plodding. It is indicative of older persons returning to an academic environment after interrupting their education. By being overly concerned about the correctness of their responses, they work slowly and carefully. This often results in their being unable to finish a timed test. It would seem unfair and unwise to base a person's ability on items that they did not reach. The only exception might be in the instance where speed was an important criterion. Since this individual completed over 80% of the test, I would argue that this person's true ability should be calculated by omitting the last 7 items. The computer program, PANAL, has an option which can automatically score tests with omitted items excluded. Using this option the ability estimate calculated by excluding omitted items is printed along with the ability estimate based on the total item set.

Since ability estimates in the Rasch model are independent of the item sets from which they are derived, it is possible to change the item sets and still have measures which are on the same scale as measures for people who took the whole test. For this individual the ability estimate excluding the omitted items would be about 3.0, 1.2 logits higher than the

ability estimate including the omitted items. This is not to say that all omitted items should be excluded from the estimation procedure, but in this case omitting the last seven, unattempted, items would produce a more accurate ability estimate.

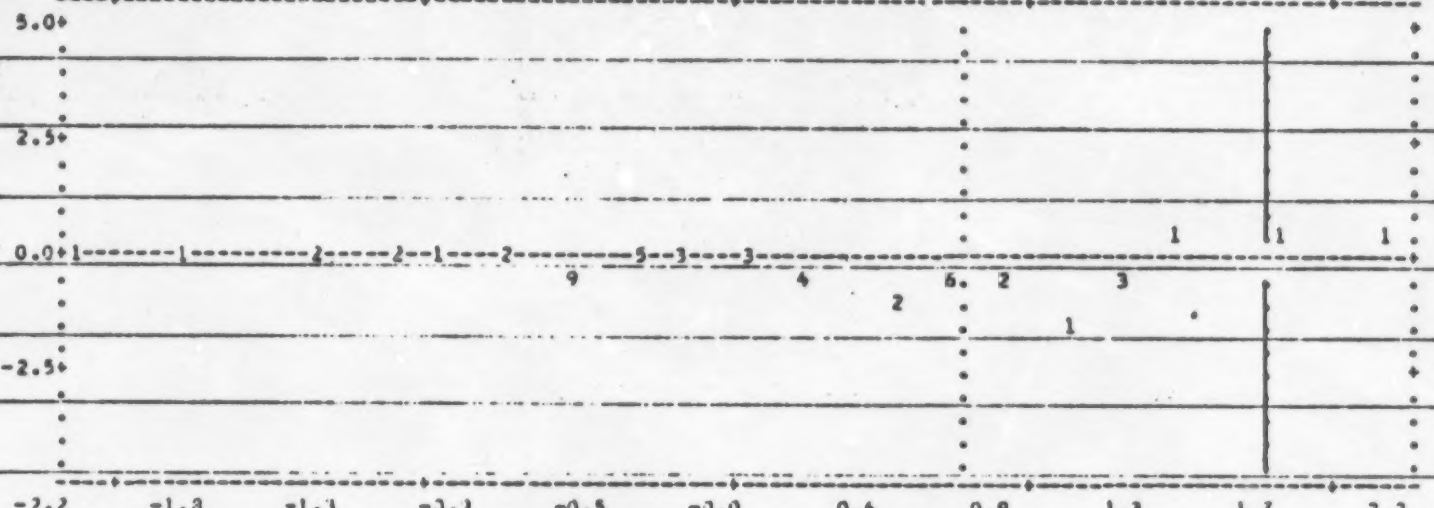
PANAL 3 — SELECT BI GREATER THAN 1.5

PERSON IDENTIFICATION 5

SUMMARY STATISTICS

MEASURE	1.76
SEM	0.39
TOTAL T	-1.0
BETWEEN SUBTEST SUBSCORES	2.4
SEQUENCE	1.4
DIFFICULTY	-0.5

MEAN RESIDUAL BY DIFFICULTY



-2.2 -1.8 -1.3 -0.9 -0.5 -0.0 0.4 0.9 1.3 1.7 2.2

ITEM DIFFICULTY IN LOGITS

NAME	COUNT	SCORE	ABILITY	S.E.	WITHIN	BETWN
SUBSCORES			1.76	0.35	-1.0	2.4
SUB1	12.	12.	9.99	9.99	9.99	-5.1
SUB2	13.	11.	2.41	0.80	-0.9	
SUB3	12.	12.	9.99	9.99	9.99	-5.6
SUB4	13.	6.	0.25	0.58	1.6	

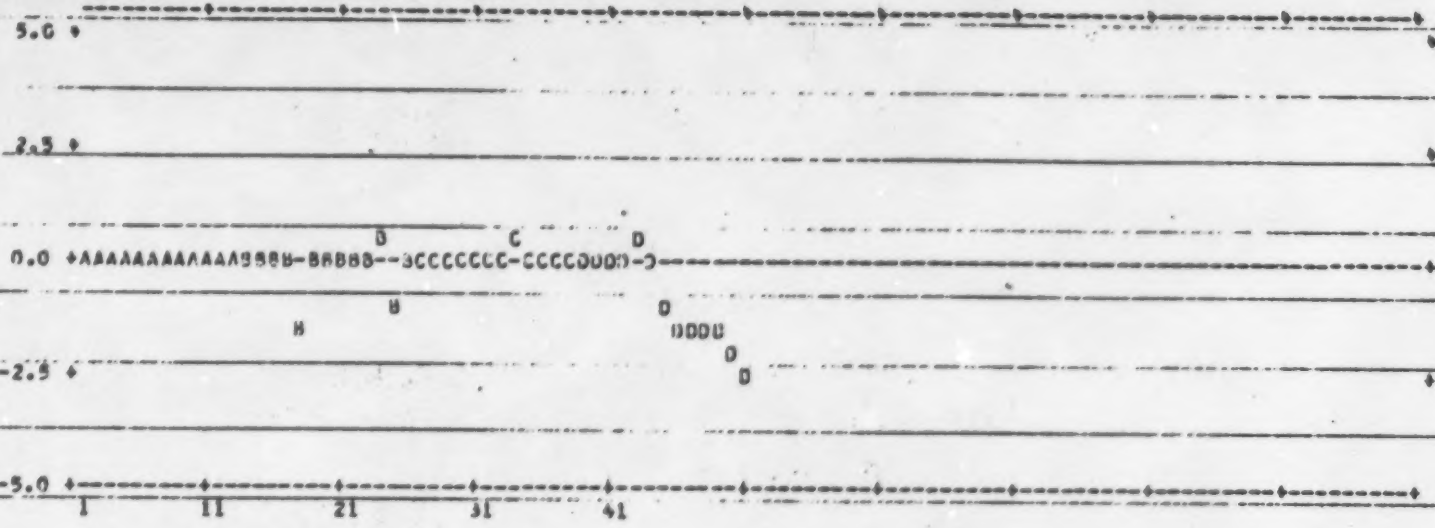
NAME	COUNT	SCORE	ABILITY	S.E.	WITHIN	BETWN
SEQUENCE			1.76	0.35	-1.0	1.4
1HLF	25.	23.	2.71	0.77	-3.1	
2HLF	25.	18.	1.20	0.47	0.8	

NAME	COUNT	SCORE	ABILITY	S.E.	WITHIN	BETWN
DIFFICULTY			1.76	0.39	-1.0	-0.5
-.75	9.	9.	9.99	9.99	9.99	-5.1
0.00	18.	17.	2.47	1.03	-1.2	
0.80	15.	11.	1.49	0.59	0.6	
2.00	8.	4.	1.41	0.72	1.1	

PANEL 3 -- SELECT RI GREATER THAN 1.5

PERSON IDENTIFICATION 5

LOG LOGIT RESIDUALS PLOTTED IN SEQUENCE ORDER



Person #6 answered 11 of the 30 items on the Elementary Algebra test correctly. This resulted in an initial ability estimate of -0.63 logits. The total test fit statistic, page A-11, is -1.2, indicating a very consistent pattern of responses. This person had all of the 6 easiest items correct, none of the 5 hardest items correct, and very few surprising responses. Again in this case there is no indication of a problem in either the total test fit statistic or in the mean residual by difficulty plot.

The subscore set analysis adds little to our understanding other than eliminating certain possibilities from consideration. Since all three of the within set fit statistics are acceptable, it is unlikely the problem is guessing, even though the raw score of 11 might lead us to originally consider this. The difficulty set analysis confirms the fact that this person tends to fit the model too well. So well in fact that the between set fit statistic is getting large. Notice, however, the pattern of ability estimates plotted on the right. A line connecting the plotted ability estimates has the opposite slope of a line connecting the points in a guessing case, see page A-1. Even though the between set fit statistic may be large, the direction of change of the ability estimates can be informative.

Looking at the residual plot in sequence order on page

A-12, we can begin to see the reason for the misfit, 12 incorrect items beginning with item #19. A check of the original responses indicated that this individual omitted the last 10 items. Two points differentiate this person from person #5. This person omitted substantially more items, 33% vs. 14%, and this person has a substantially lower original ability estimate, -0.63 vs. 1.76. Despite these two differences, we can still detect the misfit for this person. We should still be able to get a valid ability estimate on the basis of the 20 items attempted. Deleting the 10 omitted items would yield an ability estimate of -0.1 logits with a standard error of 0.4. Omitting the 10 items does not seriously affect the standard error. In a case as severe as this, it might also be advisable to recommend that this individual receive some assistance in time management in testing situations and might benefit from an instructional program with one-to-one interaction.

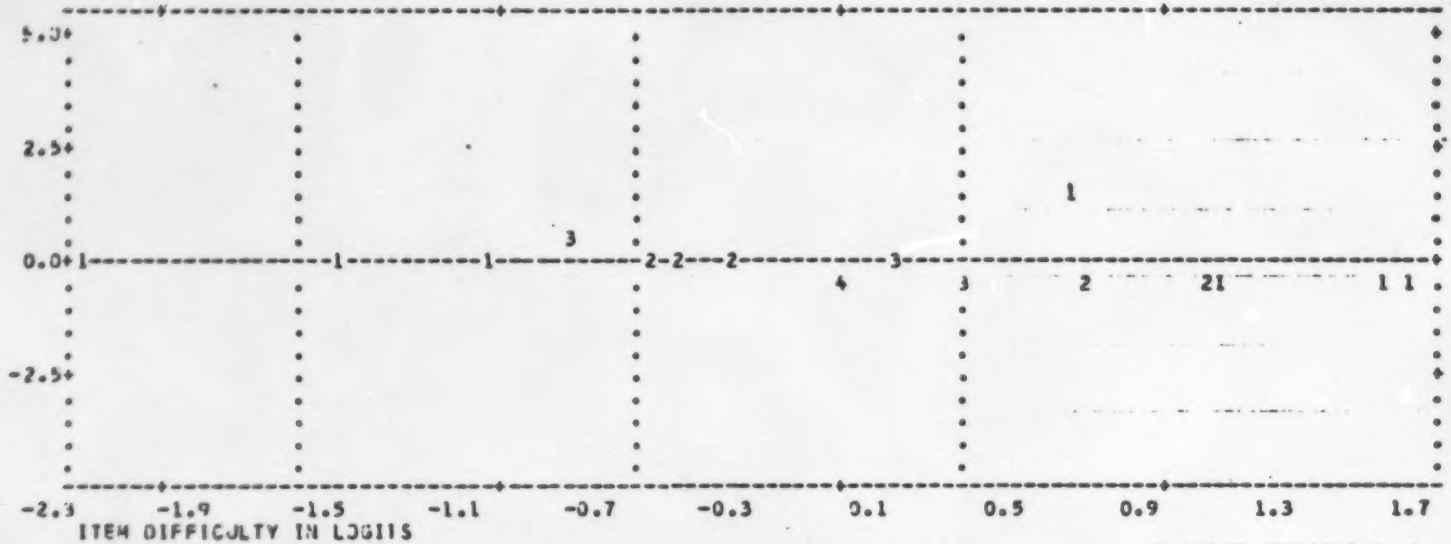
PANEL 3 -- SELECT ON GREATER THAN 1.5

PERSON IDENTIFICATION 6

SUMMARY STATISTICS

MEASURE	-0.63
S.E.	0.41
TOTAL T	-1.2
BETWEEN SUBTEST SEQUENCE	1.5
SUBSCORES	0.4
DIFFICULTY	1.0

MEAN RESIDUAL BY DIFFICULTY



NAME	COUNT	SCORE	ABILITY	S.E.	WITHIN	BETWN
SEQUENCE			-0.63	0.41	-1.2	1.5
ITEM	10.	5.	-0.33	0.70	-0.4	
2ITEM	10.	6.	0.22	0.68	0.1	
3ITEM	10.	0.	-999.99	999.99	-4.8	

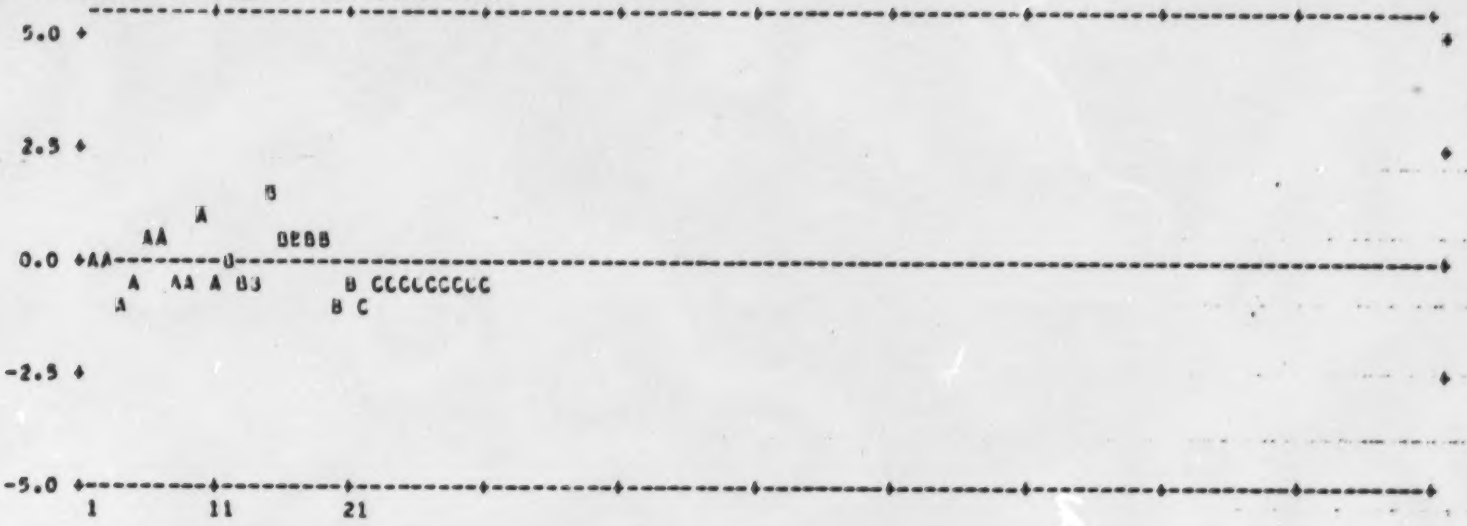
NAME	COUNT	SCORE	ABILITY	S.E.	WITHIN	BETWN
SUBSCORES			-0.63	0.41	-1.2	0.4
SUB1	18.	8.	-0.33	0.51	-0.3	
SUB2	12.	3.	-1.17	0.73	-1.3	

NAME	COUNT	SCORE	ABILITY	S.E.	WITHIN	BETWN
DIFFICULTY			-0.63	0.41	-1.2	1.0
-.75	6.	6.	999.99	999.99	-3.4	
0.00	8.	3.	-1.17	0.73	0.2	
0.00	11.	2.	-1.17	0.74	-0.1	
2.00	5.	0.	-999.99	999.99	-3.8	

PANEL 3 -- SELECT IF GREATER THAN 1.5

PERSON IDENTIFICATION 6

LOG LOGIT RESIDUALS PLOTTED IN SEQUENCE ORDER



IV D. CONTENT INTERACTION

Person #7 had a raw score of 34 on the Logical Relationships test. This person's ability estimate was 0.88. Looking at page A-13, we see that this individual's total test fit statistic was 0.4. The response pattern summarized in the mean residual by difficulty plot indicates a pattern of responses in keeping with what our model predicts should happen. With the exception of two easy items missed, the residuals are all near the 0.0 line. We should note again that if we had been looking only at a total test fit statistic. this person would not have been brought to our attention. The subscore set analysis yields a between set fit statistic of 2.7. Sub1 had a low ability estimate, -0.87. Sub2 had a mid-range ability estimate, 0.65, while sub3 and sub4 had fairly high ability estimates, 2.47 and 1.75. the within set fit statistics indicate that only the responses to sub1 are inconsistent. The sequence set analysis again yields a between set fit statistic of 2.7. The first half ability estimate is considerably lower than the second half ability estimate, over 2 logits. The first half within fit statistic also indicates some inconsistency, fit=1.6.

The difficulty set analysis yields a between set fit statistic of 1.2. This is a large value, but to a large extent this is due to the fact that this person fit the model too well. Notice that this individual answered only 1 of the 8 hard

items correctly, less than we would have expected. As a result, the between set fit is high, but it is not necessarily an indication of a problem.

The within set fit of 1.8 for the easy items, labeled -0.75, is a problem that is worth discussing. If we order these 9 items from easiest to most difficult and look at this person's responses, scored dichotomously, we have the following pattern: 110110111. This is a fairly inconsistent pattern, but it is good to remember that the within set fit statistic is calculated using the total test ability estimate. Even if this person's response pattern had been 111111100, the within set fit statistic for this person would have been over 1.0, because the two incorrect responses were 2 logits below the person's ability and these two incorrect respones, no matter where they might have been in the easy item set, would have given a large fit value.

If we look at the sequence order residual plot, we begin to get an understanding of the problem. The similarity between this individual and the previous two individuals, with startup problems, is strong. Looking only at the first, third, and fourth subscore groups, one would be inclined to classify this misfit as startup also. The high within set fit, 1.8, for sub1 tends to confirm this. There is something else going on here, though. The change in the residual plot by sequence order between sub2 and sub3, labeled B and C on page A-14, is too

pronounced. It is too coincidental that this change occurs exactly when the content changes. This person's ability based on connectives is certainly different than the ability based on either analogues or principles of organization. It would appear that the fact that this individual is less able in one, possibly two, of the content areas is the primary reason for the misfit. There is also an indication of startup. Unfortunately, because of the structure of the test, the first 12 items are from the same content area, it is difficult to distinguish between startup and a content/person interaction with the first content area. The mixing of items from the various content areas, as is the case on the Elementary Algebra test, makes it easier to separate these two types of misfit.

The possible presence of two types of misfit, startup, and content/person interaction is supported by the high within set fit, 1.8, for sub1. However, the fact that most of the disturbance in sub1 occurred in the last six items would argue against the presence of startup. The precise identification of the type of source of the misfit is unnecessary. The important point is that we recognize that in fact the total test score is not a true representation of the individual's ability. I might be inclined to accept the characterization of this problem as resulting from an unequal ability in this person, lower on sub1 and sub2, higher on sub3 and sub4. As a result of this decision, we would report only the four subscores and no total

test score.

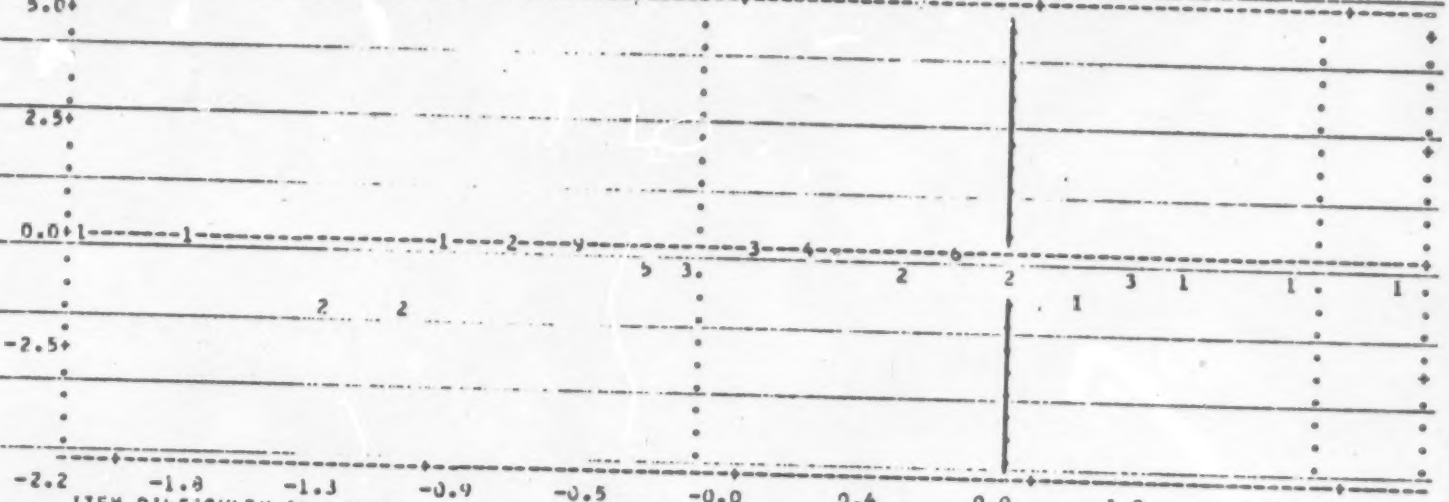
PAVAL 3 -- SELECT B1 GREATER THAN 1.9

PERSON IDENTIFICATION 7

SUMMARY STATISTICS

PLEASURE	0.88
SD4	0.33
TOTAL T	0.4
BETWEEN SUBTEST	
SUBSCORES	2.7
SEQUENCE	2.7
DIFFICULTY	1.2

MEAN RESIDUAL BY DIFFICULTY



NAME	COUNT	SCORE	ABILITY	S.E.	WITHIN	BETWN
SUBSCORES			0.88	0.33	0.4	2.7
SUB1	12.	6.	-0.82	0.61	1.8	
SUB2	13.	7.	0.65	0.60	-0.1	
SUB3	12.	11.	2.47	1.07	-3.0	
SUB4	13.	10.	1.75	0.69	-1.3	

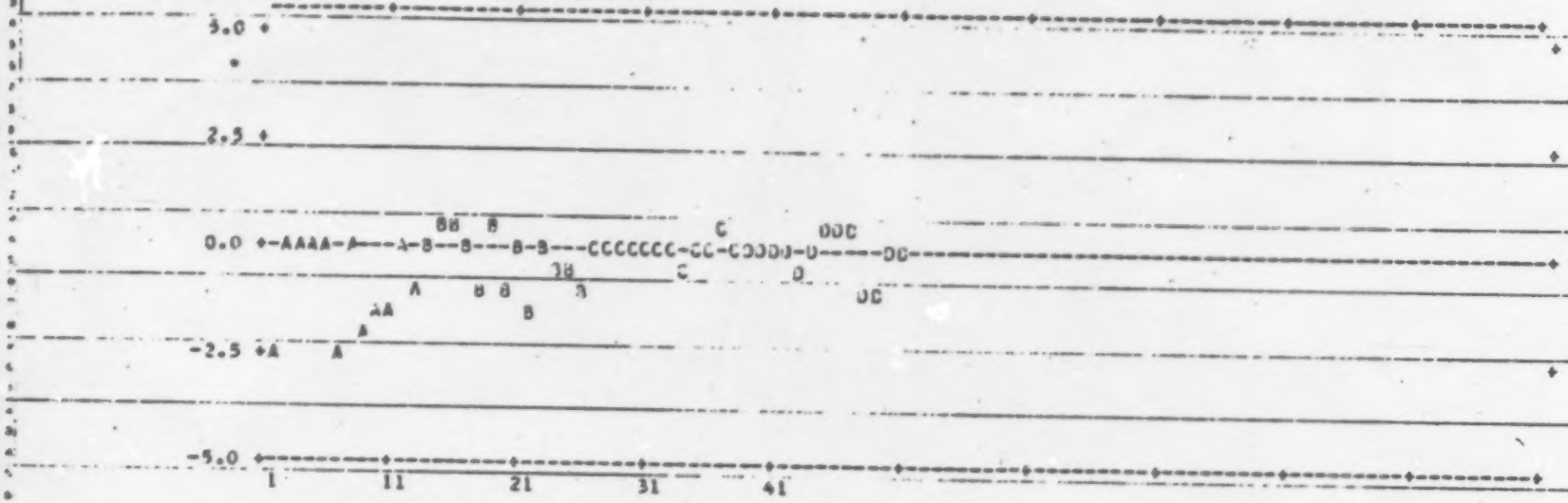
NAME	COUNT	SCORE	ABILITY	S.E.	WITHIN	BETWN
SEQUENCE			0.88	0.33	0.4	2.7
1HLF	25.	13.	-0.06	0.44	1.6	
2HLF	25.	21.	1.69	0.57	-3.0	

NAME	COUNT	SCORE	ABILITY	S.E.	WITHIN	BETWN
DIFFICULTY			0.88	0.33	0.4	1.2
-2.5	9.	7.	0.03	0.60	1.8	
0.00	13.	14.	0.46	0.57	0.3	
0.80	15.	12.	1.87	0.65	-0.9	
2.00	5.	1.	-0.55	0.57	-1.0	

PANEL 3 -- SELECT B1 GREATER THAN 1.5

PERSON IDENTIFICATION 7

LOG LOGIT RESIDUALS PLOTTED IN SEQUENCE ORDER



Person #8 had a raw score of 12 on the Elementary Algebra test. This resulted in an ability estimate of -0.46. The fit statistics on page A-15 indicate that the total test fit statistic for this person was 2.4. The mean residual by difficulty plot shows the typical misfit residual pattern, with this individual missing several easy items and getting several difficult items correct. The subscore set analysis shows a between set fit of 23.6, a very large number. There is a 4.4 logit difference between this person's ability estimated from the solving equation items and from the items involving algebraic operations. The large within set fit, 3.1, for sub2 seems to be more the result of consistently answering the difficult items in sub2 correctly based on the overall ability estimate of -0.46 than from an inconsistency of responses within this set. the sequence set analysis yielded a between set fit of 8.2. Two of the within set fit statistics, 2ten and 3ten, are high, 1.5 and 2.6. The difficulty set analysis yielded a between set fit of 7.0 and two large within set fit statistics, high middle, 1.3, and hard, 2.0.

The effect of the content interaction can clearly be seen in the residual plot in sequence order. In this plot, B represents sub2 and A represents sub1. It is interesting how an extreme case of a single type of misfit can have a profound effect on all of the misfit statistics. The sequence set analysis is affected because 8 of the 10 items in 2ten are from

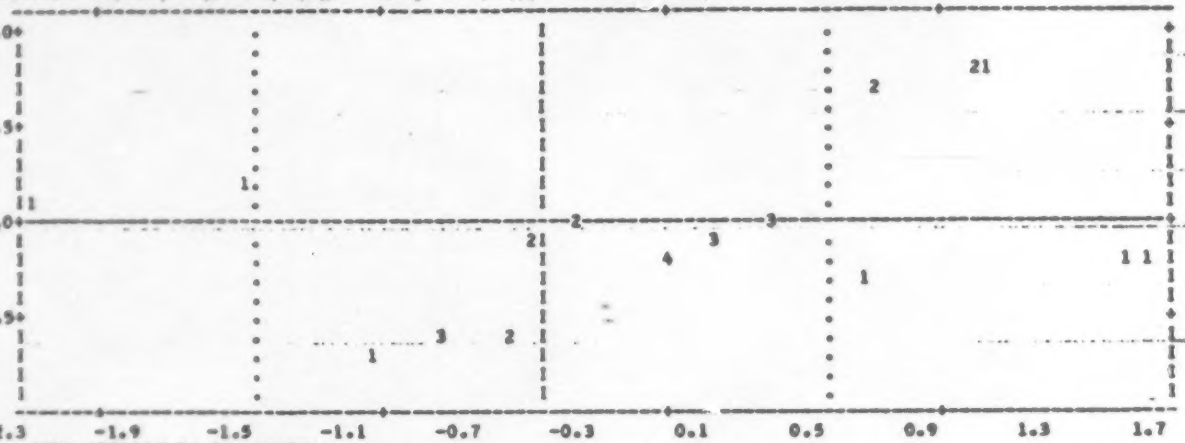
subi, resulting in a low ability estimate. Likewise, 7 of the 10 items in 3 ten were from sub2, resulting in a higher ability estimate. All of the within set fit statistics were affected by this individual's ability to answer difficult items about algebraic operations correctly. There would appear to be little to be gained by retesting this individual; although confirming this pattern with a second test might be informative, there is no reason to believe that it would change drastically. One thing is certain, -0.46 does not truly represent this person's ability. Two separate scores must be reported here. The ultimate placement recommendation here must be aided by an in-person interview.

NJCbSPT ELEMENTARY ALGEBRA TOTAL T GREATER THAN 1.5

PERSON IDENTIFICATION 8

SUMMARY STATISTICS	
MEASURE	-0.46
SEM	0.40
TOTAL T	2.4
BETWEEN SUBTEST SUBSCORES	23.6
SEQUENCE	8.2
DIFFICULTY	7.0

MEAN RESIDUAL BY DIFFICULTY



NAME	COUNT	SCORE	ABILITY	S.E.	WTHIN	BETWN
SUBSCORES			-0.46	0.40	2.4	23.6
			-0.46	0.40		
SUB1	18.	2.	-2.38	0.77	0.1	
SUB2	12.	10.	1.90	0.80	3.1	

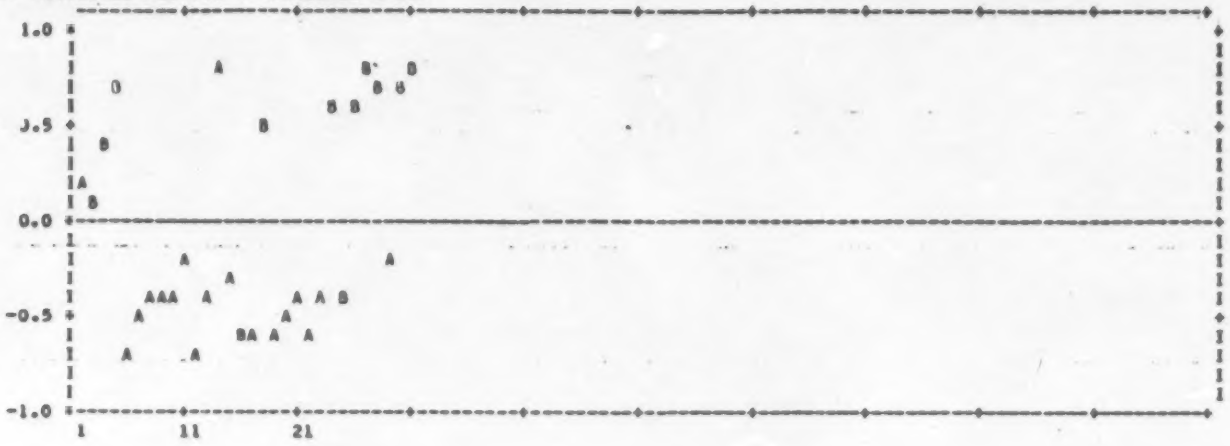
NAME	COUNT	SCORE	ABILITY	S.E.	WTHIN	BETWN
SEQUENCE			-0.46	0.40	2.4	8.2
			-0.46	0.40		
1TEM	10.	4.	-0.83	0.71	-0.3	
2TEM	13.	2.	-1.72	0.81	1.5	
3TEM	10.	6.	1.00	0.67	2.6	

NAME	COUNT	SCORE	ABILITY	S.E.	WTHIN	BETWN
DIFFICULTY			-0.46	0.40	2.4	7.3
			-0.46	0.40		
-.75	6.	2.	-1.96	0.90	0.8	
0.00	8.	2.	-1.46	0.82	-9.7	
0.83	11.	5.	0.14	0.61	1.3	
2.00	5.	3.	1.72	0.92	2.0	

NJCBSPT ELEMENTARY ALGEBRA TOTAL T GREATER THAN 1.5

PERSON IDENTIFICATION 8

LOG LOGIT RESIDUALS PLOTTED IN SEQUENCE ORDER



IV E. SLOPPINESS (CARELESSNESS)

Person #9 answered 46 of the 50 items on the Logical Relationships test correctly. The ability estimate for this individual was 2.77. The total test fit statistic, found on page A-17, was 3.7. The mean residual plot by difficulty indicates that this individual missed 3 items which were at least 3 logits below his ability. The probability of this person answering each of these items incorrectly is less than 0.04. The subscores between set fit was 0.9. Contrasting the within-set fit for sub1 and sub2 gives a good indication of the sensitivity of this fit statistic. In both cases, this individual missed one item in each of the sets. In sub1 the incorrect item was 3.9 logits below the individual's ability. In sub2 the incorrect item was 1.4 logits below the individual's ability. An incorrect response to the second item is 10 times as likely, P=0.20 vs. P=0.02, than an incorrect response to the first, and the within set fit statistics reflect this, 4.4 for sub1 vs. -1.5 for sub2.

The sequence set analysis indicates that there is little difference between the ability estimated on the first half and the second half, but the large within set fit statistics indicate that there are some improbable responses on both halves of the test. The difficulty set analysis indicates that there is a difference between the ability estimates based on the 4 item difficulty groupings. The 2.5 logit difference is

expressed by a between set fit of 1.7. Again, notice the sensitivity of the within set fit statistic to the difficulty of the item answered incorrectly; missing one easy item resulted in a within set fit of 4.6 for the easy items, while missing one item in the hard item set resulted in a within set fit of -0.2.

The residual plot in sequence order, page A-18, is characteristic of sloppiness. The easy items that were missed do not appear to have any systematic basis, although there may be some underlying commonality which escapes detection without careful analysis of the items involved. A strong argument can be made for the fact that the ability estimate of 2.77 is in fact an underestimate of this person's true ability. Three of the four incorrect items are at least 3 logits away from this person's ability. These responses contain 1/5 to 1/10 of the information that items centered on the person's ability would contain. Deleting these items would have little impact on the amount of information contained in this test and would give a better estimate of this person's ability. If these 3 items were deleted from the scoring procedure, this individual's ability would be approximately 3.8 logits, an increase of 1 logit.

An important question, "is sloppiness forgivable?", clearly must be considered before a correction such as the one suggested above is implemented. In most cases, this decision must be made in the context of the testing situation and the

nature of the decisions which prompted the testing. It is, therefore, difficult to give a universal answer, but the question should be considered.

PANAL 3 -- SELECT T GREATER THAN 1.5

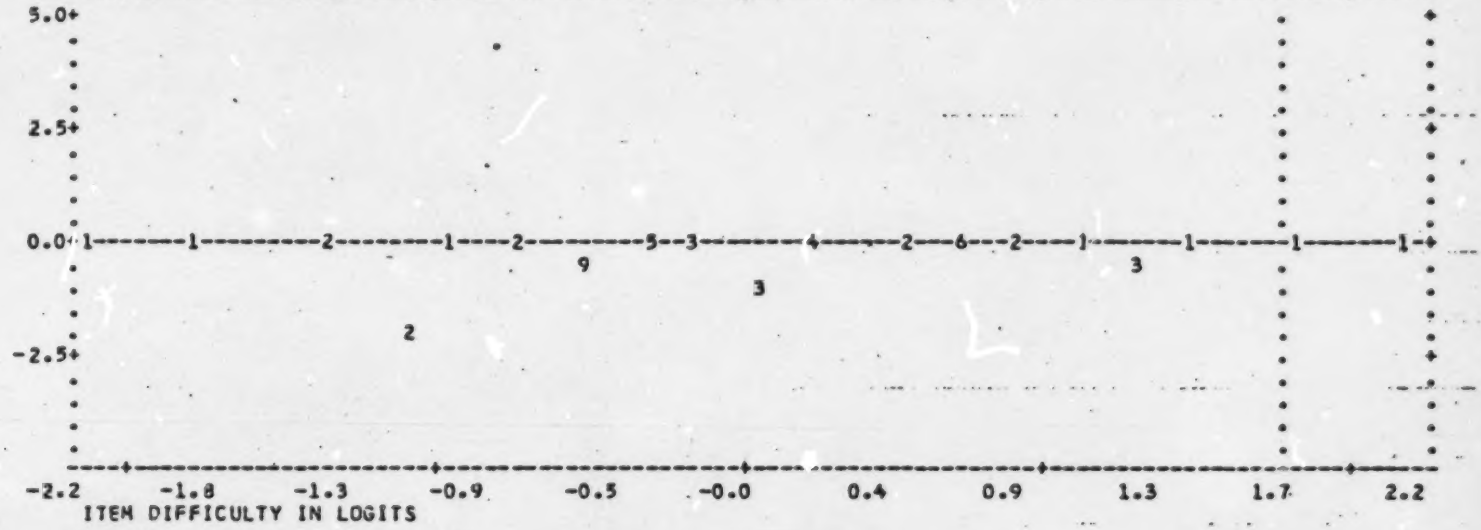
PERSON IDENTIFICATION 9

SUMMARY STATISTICS

MEASURE 2.77
SEM 0.53
TOTAL T 3.7

BETWEEN SUBTEST
SUBSCORES 0.9
SEQUENCE -1.2
DIFFICULTY 1.7

MEAN RESIDUAL BY DIFFICULTY



NAME	COUNT	SCORE	ABILITY	S.E.	MTHIN	BETWN
SUBSCORES			2.77	0.53	3.7	0.9
SUB1	12.	11.	1.76	1.06	4.4	
SUB2	13.	12.	3.25	1.07	-1.5	
SUB3	12.	10.	1.62	0.81	3.9	
SUB4	13.	13.	9.99	99.99	9.90	-6.5

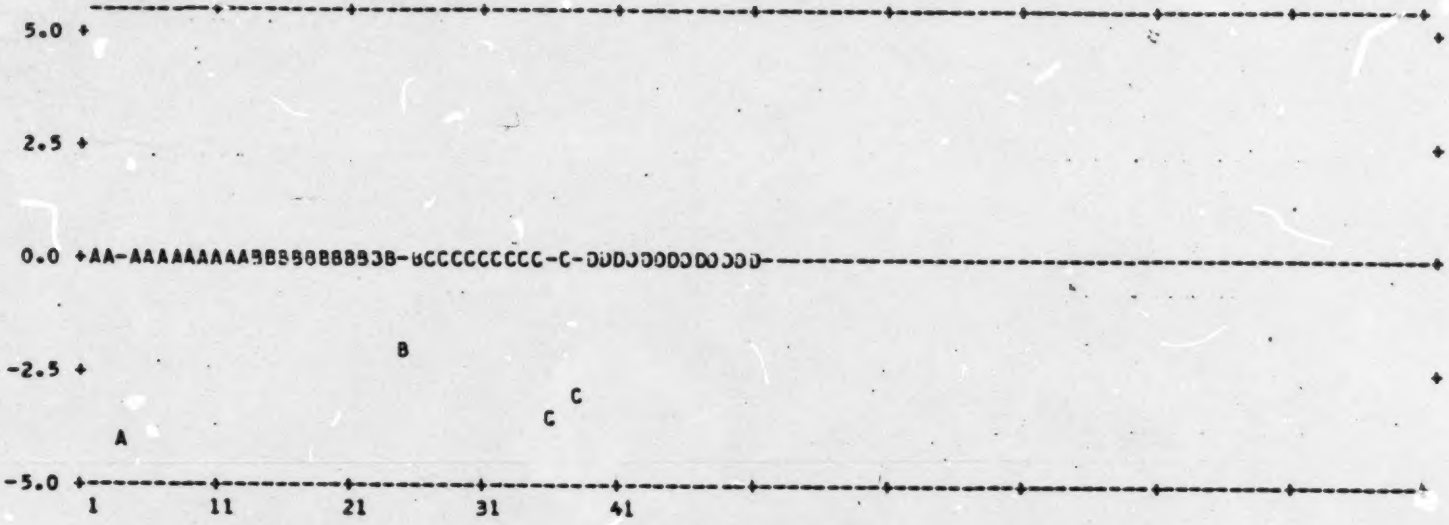
NAME	COUNT	SCORE	ABILITY	S.E.	MTHIN	BETWN
SEQUENCE			2.77	0.53	3.7	-1.2
1HLF	25.	23.	2.71	0.77	3.2	
2HLF	25.	23.	2.83	0.76	2.4	

NAME	COUNT	SCORE	ABILITY	S.E.	MTHIN	BETWN
DIFFICULTY			2.77	0.53	3.7	1.7
-.75	9.	8.	0.87	0.99	4.6	
0.00	18.	16.	1.71	0.75	3.3	
0.80	15.	15.	9.99	99.99	9.90	-7.5
2.00	8.	7.	3.41	1.06	-0.2	

PANEL 3 -- SELECT T GREATER THAN 1.5

PERSON IDENTIFICATION 9

LOG LOGIT RESIDUALS PLOTTED IN SEQUENCE ORDER



Person #10 answered 28 of the 30 items on the Elementary Algebra test correctly. This resulted in an ability estimate of 2.94 logits. The fit statistic on page A-20 indicates a total test fit of 10.4. The reason is clear on the mean residual plot. This individual missed the easiest item on the test, an item which is over 5 logits below this person's ability. The probability of this occurring is less than 0.01.

Looking at the 3 set analyses, we can see that the within set fit for any set containing that item was severely impacted. The sequence residual plot indicates that this was the first item on the test. We might classify this as either sloppiness or startup, but in either event, this item should be omitted from the person's ability estimate. With this item omitted, the ability estimate for this person should rise to about 3.7, an increase of 0.76 logits.

PANAL 3 -- SELECT T GREATER THAN 1.5

PERSON IDENTIFICATION 20

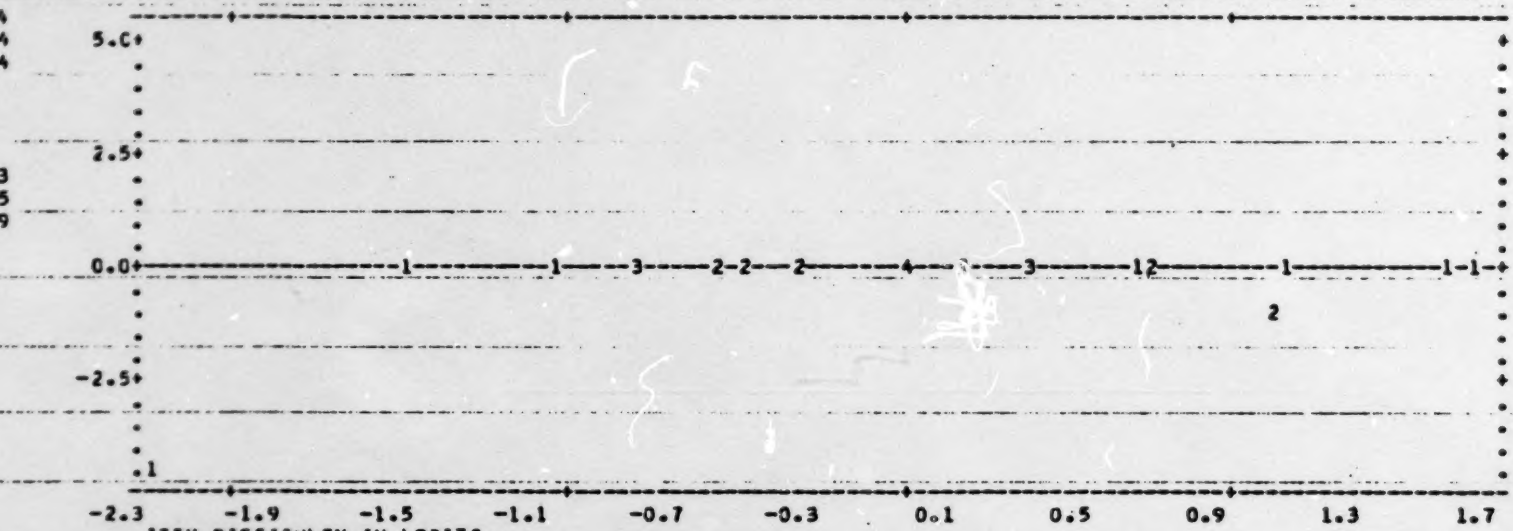
SUMMARY STATISTICS

MEASURE	2.94
SEM	0.74
TOTAL T	10.4

BETWEEN SUBTEST

SEQUENCE	-0.3
SUBSCORES	1.5
DIFFICULTY	1.9

MEAN RESIDUAL BY DIFFICULTY



NAME	COUNT	SCORE	ABILITY	S.E.	MTHIN	BETWN
SEQUENCE			2.94	0.74	10.4	-0.3
1TEN	10.	9.	2.23	1.11	11.5	
2TEN	10.	10.	9999.90	9999.90	-5.9	
3TEN	10.	9.	2.90	1.07	-0.2	

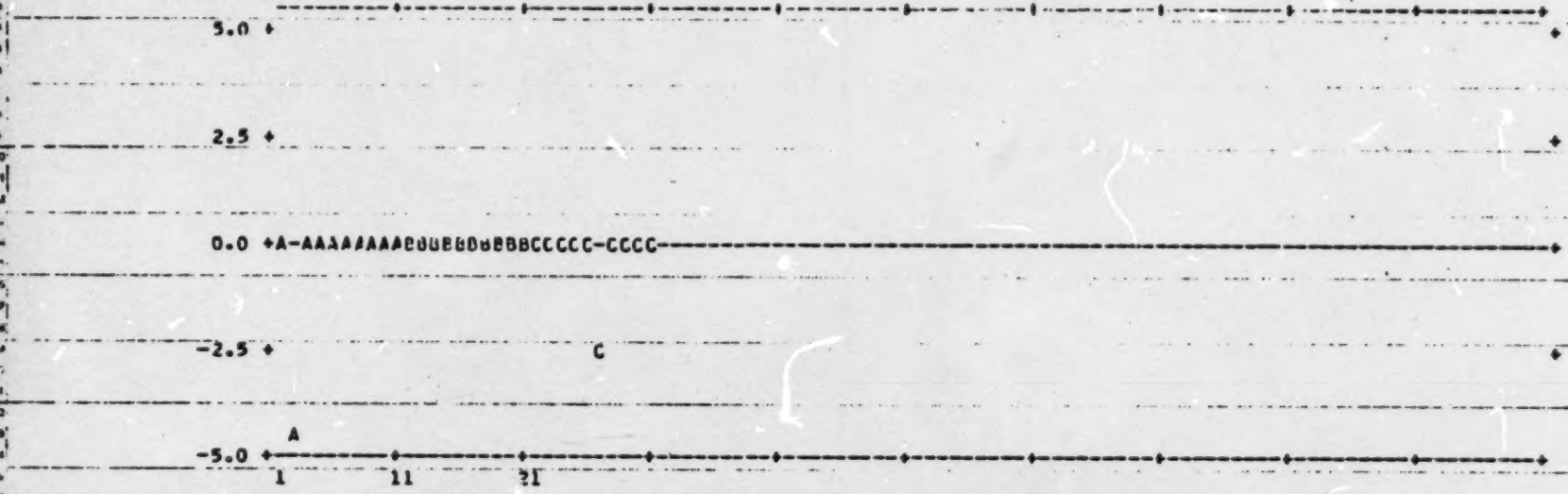
NAME	COUNT	SCORE	ABILITY	S.E.	MTHIN	BETWN
SUBSCORES			2.94	0.74	10.4	1.5
SUB1	18.	18.	9999.90	9999.90	-7.9	
SUB2	12.	10.	1.90	0.80	11.6	

NAME	COUNT	SCORE	ABILITY	S.E.	MTHIN	BETWN
DIFFICULTY			2.94	0.74	10.4	1.9
-.75	6.	5.	0.47	1.07	11.3	
0.00	8.	8.	9999.90	9999.90	-5.4	
0.80	11.	11.	9999.90	9999.90	-6.3	
2.00	5.	4.	2.72	1.12	1.2	

PANAL 3 -- SELECT T GREATER THAN 1.5

PERSON IDENTIFICATION 10

LOG LOGIT RESIDUALS PLOTTED IN SEQUENCE ORDER



IV F. REANALYSIS

The first ten examples were chosen to clearly illustrate the five characteristic types of misfit. In many instances, the detection and understanding of misfit is more difficult. Person #11 is such an example. In this instance, the original analysis had a very large sequence between-set fit statistic. This is characteristic of plodding, so the raw response vector was examined. This examination indicated that this individual had omitted the last 13 items. A reanalysis of the response vector using the omit option was completed. The omit option will exclude omitted items from the estimation and analysis procedure.

The results of this analysis are on pages A-21 and A-22. By omitting the 13 items at the end of the test and 1 earlier item, this person's ability estimate rose from -1.59 to -0.91. Once these items were removed, however, the true character of these response became evident. The total test fit statistic is now 1.6 and the difficulty between-set fit-statistic rose to 5.4. As a result of reanalysis, what had originally looked like a case of plodding could clearly be identified as guessing. It is only speculation why a person would guess for half the test and then give up, but it is obvious that no matter what the cause, this test result cannot be considered a valid score and this individual must be retested.

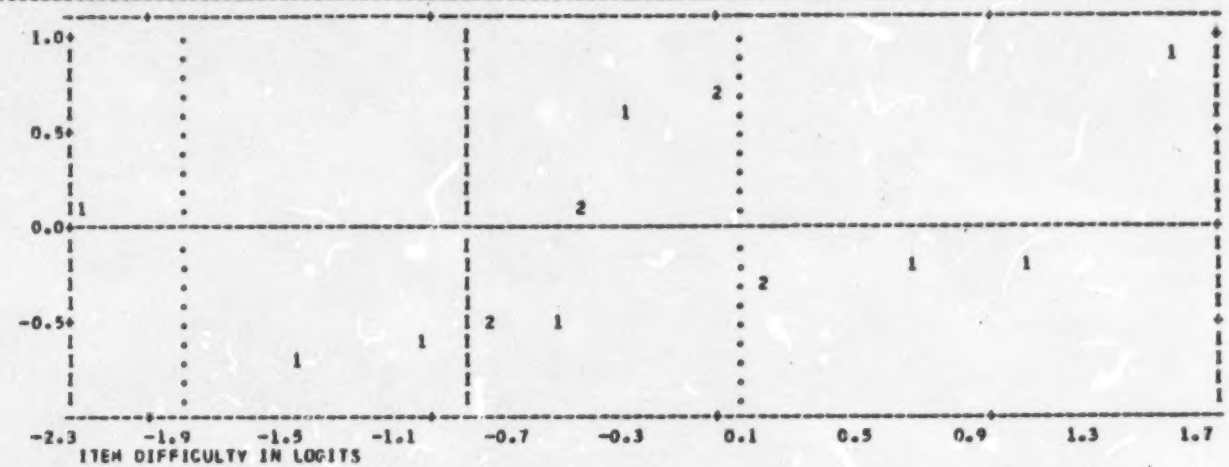
NJCBSPT ELEMENTARY ALGEBRA TOTAL T GREATER THAN 1.5

PERSON IDENTIFICATION 11 PAGE A-21

SUMMARY STATISTICS

MEASURE	-1.59
SEM	0.48
TOTAL T	1.6
SKIPPING UNITS	
MEASURE	-0.91
SEM	0.56
BETWEEN SUBTEST	
SUBSCORES	-1.8
SEQUENCE	-7.8
DIFFICULTY	5.4

MEAN RESIDUAL BY DIFFICULTY



ITEM DIFFICULTY IN LOGITS

NAME	COUNT	SCORE	ABILITY	S.E.	MTHIN	BETWN
SUBSCORES			-1.59	0.48	1.6	-1.8
			-0.91	0.56		
SUB1	12.	3.	-1.38	0.71	1.6	
SUB2	4.	3.	0.33	1.19	0.5	

NAME	COUNT	SCORE	ABILITY	S.E.	MTHIN	BETWN
SEQUENCE			-1.59	0.48	1.6	-7.8
			-0.91	0.56		
ITEM	9.	4.	-0.70	0.74	1.5	
ZTEM	7.	2.	-1.18	0.87	0.6	

NAME	COUNT	SCORE	ABILITY	S.E.	MTHIN	BETWN
DIFFICULTY			-1.59	0.48	1.6	5.4
			-0.91	0.56		
-.75	5.	1.	-2.40	1.15	0.4	
0.00	6.	4.	0.36	0.97	1.9	
0.80	3.	3.	**********		-1.1	
2.00	2.	1.	1.33	1.43	1.7	

SUBSET ELEMENTARY ALGEBRA TOTAL T GREATER THAN 1.5

PERSON IDENTIFICATION 11

LOG LOGIT RESIDUALS PLOTTED IN SEQUENCE ORDER

IV G. RANDOM RESPONSES

One of the problems in dealing with guessing when it occurs over the entire test is the fact that they are random responses. The fact that they are random means for the most part the individual will guess both easy and hard items correctly and misfit. Occasionally an individual will guess only hard items correctly and misfit, but it is equally probable that an individual will guess only easy items correctly and fit.

To explore this possibility, we created a series of totally random responses to the Elementary Algebra test (P=0.25). To illustrate the point that random respones sometimes fit the model, look at pages A-23 and A-24. This individual (?) answered 11 of the 30 items on the Elementary Algebra test correctly. The total test fit statistic was 0.6. The mean residual by difficulty plot looks reasonable. All of the between set and within set fit statistics are acceptable. One might question the 2.4 logit difference between the abilities estimated for the low middle and high middle item difficulty groups, but this difference does not seem to bother the fit statistic. If this were a real person who had guessed, this person would have been placed into a second level algebra course at the college where this test is employed when in fact they were random responses.

The point is that when dealing with truly random

response, no system of screening can detect every case, and in fact it is almost a certainty that there would be some invalid results reported even if the system of screening discussed here were implemented. Fortunately, in real data truly random responses for an entire test are rare. Most of the disturbances to measurement come from systematic effects like the ones illustrated. The identification of systematic effects can be undertaken quite successfully. An ongoing program of measurement, quality control will detect sizable quantities of misfit, and in most cases provide the user with adequate clues as to the source of misfit to enable revised ability estimates to be generated. In the rare cases of truly random responses, only a portion will go undetected. In the next section, we shall look more closely at the frequency with which these types of misfit occur in testing situations.

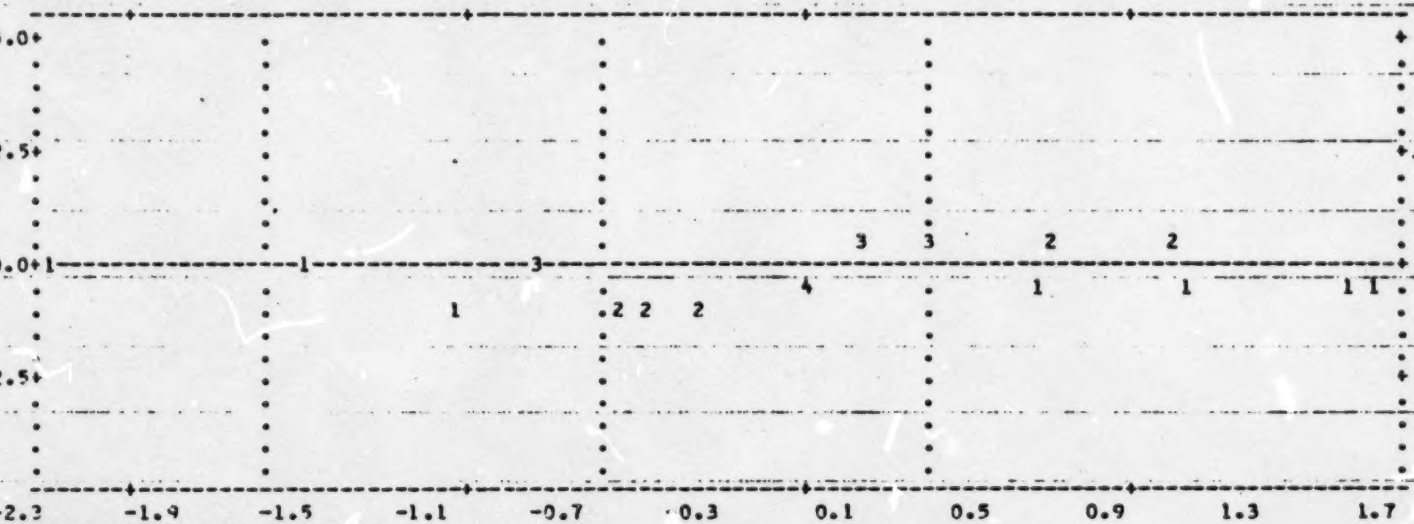
PANEL 3--SELECT ALL CASES

PERSON IDENTIFICATION 12

SUMMARY STATISTICS

MEASURE	-0.63
SEM	0.41
TOTAL T	0.6
BETWEEN SUBTEST	
SEQUENCE	-0.4
SUBSCORES	-1.2
DIFFICULTY	1.0

MEAN RESIDUAL BY DIFFICULTY



NAME	COUNT	SCORE	ABILITY	S.E.	WTHIN	BETWN
SEQUENCE			-0.63	0.41	0.6	-0.4
ITEM	10.	5.	-0.33	0.70	0.6	
2TEM	10.	3.	-1.14	0.71	-0.1	
3TEM	10.	3.	-0.36	0.72	1.2	

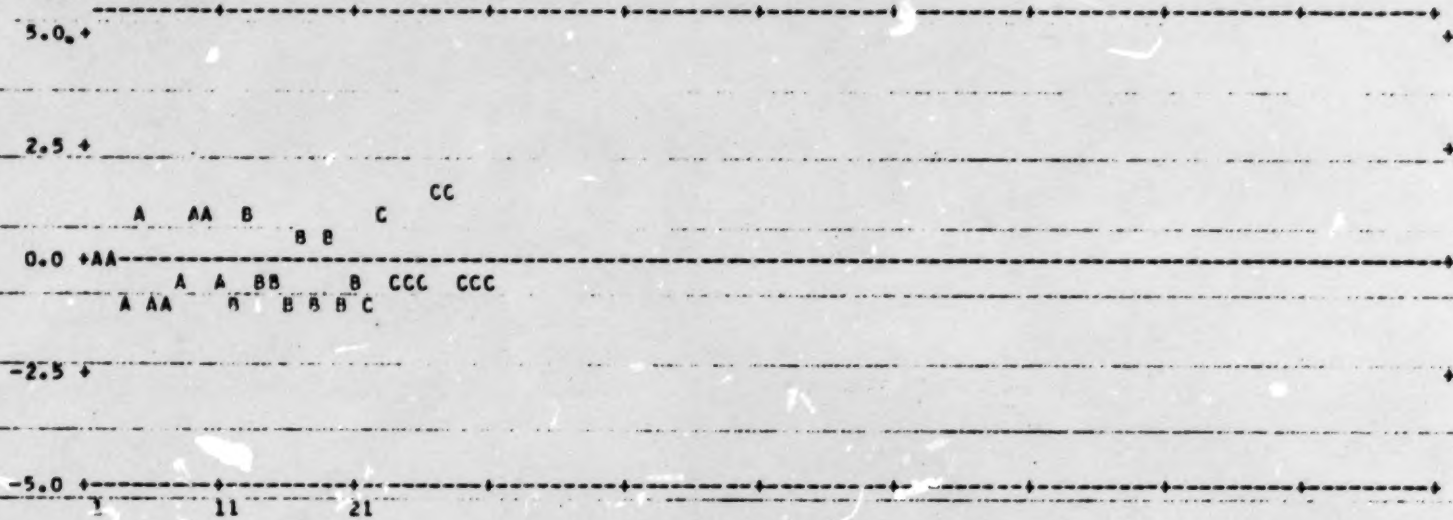
NAME	COUNT	SCORE	ABILITY	S.E.	WTHIN	BETWN
SUBSCORES			-0.63	0.41	0.6	-1.2
SUB1	18.	7.	-0.59	0.52	0.3	
SUB2	12.	4.	-0.68	0.67	1.2	

NAME	COUNT	SCORE	ABILITY	S.E.	WTHIN	BETWN
DIFFICULTY			-0.63	0.41	0.6	1.0
-.75	6.	4.	-0.48	0.89	0.2	
0.00	8.	1.	-2.31	1.07	-0.7	
0.80	11.	5.	0.14	0.61	1.0	
2.00	5.	1.	-0.10	1.08	0.3	

PANAL 3--SELECT ALL CASES

PERSON IDENTIFICATION 12

LOG LOGIT RESIDUALS PLOTTED IN SEQUENCE ORDER



V. A STUDY OF THE FREQUENCY OF PERSON DISTURBANCES IN MEASUREMENT

The psychometric model proposed by Georg Rasch is the result of his attempt to define the conditions necessary to obtain objective measurement. The Rasch Model differs from other item response theory models in its focus. This model, rather than seeking to include parameters which tend to mold the model to observations, seeks to identify in the observations those characteristics which are undesirable in objective measurement. As a result, although the Rasch Model is mathematically equivalent to the one-parameter logistic model, there is an underlying philosophic difference in the application of the Rasch model and models which seek to describe data.

Rasch's notion of objective measurement, which sought to exclude from measurement the confounding properties of testing situations that people often attribute to items, such as guessing, and item characteristic curves with different slopes, leads quite naturally to the investigation of person response patterns. The investigation of these response patterns is in fact the investigation of the fit of the person's performance to the psychometric model. There are numerous types of these item-person interactions which are deterrents to measurements. These include, but are not limited to, guessing, practice

effect, "plodding", item-type interactions, and "sloppiness", or "boredom". It is the identification of the frequency of these interactions which will be the primary focus of this section. The process of measurement as defined by Wright and Stone (1979) is as folllows:

> "Four steps must be taken to use a test to measure a person. First, we must work out a clear idea of the variable we intend to make measures on. Second, we must construct items which are believable realizations of this idea and which elicit signs of it in the behavior of the persons we want to measure. Third, we must demonstrate that these items when taken by suitable persons can lead to results that are consistent with our intentions. Finally, before we can use any person's score as a basis for their (sic) measure, we must determine whether or not their (sic) particular pattern of responses is, in fact, consistent with our expectations."

Much of the work in traditional test theory and many of the applications of the various item-response models have been concerned with the first three requirements, the construction and analysis of items and groups of items. Little, if any, attention has been given to the final requirement, the analysis of the individual person's pattern of responses. The Rasch Model, which limits measurement to two basic concepts, the ability of the person taking an item and the difficulty of the item, provides an excellent framework for conducting an analysis of an individual's pattern of responses or "person fit". Using the Rasch Model, we are able to examine the individual's response pattern for systematic discrepancies by removing the effects of the model from the data. It is the

ability to determine the probability of a correct response for any individual on any item given the individual's ability and an item pool calibrated with the Rasch Model that allows one to generate residuals (departures from expected values) by matching these probabilities with actual results.

Mead (1979) showed the effect of various item-person interactions on item characteristic curves. The effects of guessing, speed (practice) and bias were shown through a study of the residuals derived from the difference between the observed and predicted outcomes. Mead and Wright (1977) used this residual to study the effect of inappropriate test-taking behaviors on item characteristic curves. They conclude that fluctuations in item characteristic curves are often due not to a problem with the item but rather to inappropriate test-taking behaviors. They recommend procedures for removing the effect of this behavior from both the item characteristic curve and the individual ability estimates.

Wright (1977) continued the discussion of person fit and illustrated the properties of guessing and "sleeping" by plotting the residuals against the estimated differences between person abilities and item difficulties (b-d). In the guessing case, as the (b-d) difference becomes negative, the probability of a correct response becomes small; a correct response due to guessing would, therefore, have a large residual. This residual would become larger as the items become

more difficult, giving the plot of the residuals for correct responses an upward trend.

Wright and Stone (1979) conclude:

> Therefore, if we are serious in our attempt to measure, we must examine every application to see how well each set of respones corresponds to our model expectations. We must evaluate not only the plausibility of the sample of persons' responses, but also the plausibility of each person's responses to the set of items in his test. To do this we must examine the responses of each person to each item to determine whether it is consistent with the general pattern of responses observed.

It was this viewpoint which led Mead (1978) to write a computer program "PANAL" which was used to calculate and display the various aspects of a person's fit statistics for tests composed of previously calibrated items.

The purpose of this study is to analyze person fit statistics on various college placement tests, examine the effect of item difficulties on these statistics, improve the techniques for identifying misfitting individuals, and discuss techniques for improving the measure of the abilities for these misfitting individuals.

The placement test used for this study was the New Jersey College Basic Skills Placement Test (NJCBSPT). This test was developed in response to a New Jersey Board of Higher Education mandate that the basic skills of all freshmen entering public two-year and four-year colleges in the State of New Jersey be assessed and that the results of this assessment be used for college placement (but not admissions). The NJCBSPT is

published by Educational Testing Services and consists of five objective tests and one twenty minute essay. There are three thirty-minute objective English tests -- Reading Comprehension, Sentence Structure, and Logical Relations. There are two thirty-five-minute objective mathematics tests -- Computation and Elementary Algebra. The English tests were developed from the item pool used by Educational Testing Service to build the California State University and Colleges English Placement Test. The mathematics tests were developed from the item pool used to construct the Descriptive Test of Mathematics Skills.

The initial form of this test (Form 3AJP) 3 was administered to 822 incoming freshmen, at Mercer County Community College, during the Summer of 1978. The results of the test analysis (see Table I) indicate that one of the five versions (Elementary Algebra) was much more difficult (mean = 37% correct) than the other four tests (average mean = 71% correct). This difference was also evident in the State-wide population (N = 42,950) where the mean on the Elementary Algebra test was 47% correct.

The item response strings for 822 incoming freshmen were processed by BICAL (Wright and Mead, 1978), the Rasch Model Calibration Program, to produce item difficulties for the 185 items, and five separate ability estimates (one for each of the five objective tests) for the 822 freshmen. The item difficulties, person ability estimates, and the item response

strings were then used as input to PANAL program to calculate the individual fit statistics.

The results reported in this paper are the results of analyzing two of the tests in the NJCBSPT battery. The Logical Relationships Test and the Elementary Algebra Test were chosen to test the effect of differing average item difficulties on the distribution of fit statistics and the frequency of person misfit. A third sample of randomly generated responses were analyzed to test the frequency of detecting existing misfit through this analysis as well as to determine a general residual pattern for guessing against which the two real samples were compared.

Wright and Stone (1979) provide a review of the process for developing fit statistics for the Rasch Model. A brief overview of this development was presented in Section II.

Four fit-statistics were calculated for each indvidual's responses on these two tests. Summing the squared residuals over all the items on a particular test produced the first of the fit statistics used in this study, the "total fit". Three between-group fit statistics were also calculated. These are user-defined groups of items selected to highlight the various types of misfit discussed in Section III. Since ability estimates in the Rasch Model are "item freed", it is possible to estimate an individual's ability on any subset of the total items and compare this subset ability estimate to the total

test ability estimate. On the Elementary Algebra test, for example, the items were divided into four groups on the basis of the estimated item difficulty. On the basis of these four groups (easy, low middle, high middle, most difficult), four independent estimates of a person's ability were obtained. From these squared residuals, within group fit statistics were also generated for each group. These statistics can be evaluated like t-distributions with mean of 0 and a standard deviation of 1 as the result of applying the cube-root transformation discussed in Section III.

Three user-defined groupings were chosen for this study. They are: (see Table 2)

Sequence - Dividing the test into sequential groups based on the order of item presentation.

Subscores - Dividing the items according to item content.

Difficulty - Dividing the items into groups based on item difficulty.

These groups were chosen to highlight particular types of individual misfit. High values of the between-group fit statistic on items grouped by difficulty usually illustrates guessing. High values of the between-group fit statistic on items grouped by sequence usually indicates "plodding" (working too slowly to finish the test) or "startup" (having difficulty on early items). Additionally, in cases where items are grouped by item type or roughly arranged in item difficulty order, this fit statistic can often detect item type-person interactions or

guessing. High values of the between-group fit statistic on items grouped by content usually indicate a person-content interaction.

Table 3 shows the mean values of the fit statistic calculated for the three analyses. The large difference in the average difficulty of the Elementary Algebra Test and the Logical Relationships Test appears to account for the fact that the Logical Relationships Test fits the model slightly better than does the Elementary Algebra Test on all four of the fit statistics. The analysis of the randomly generated responses indicates a very high mean value for the total test and difficulty sub-groups fit statistics. Since both of these statistics are extremely sensitive to guessing, the result confirms our expectations and should lend support to the assumption that these test-of-fit statistics are sensitive to test-taking behavior not included in the model.

Further analysis of the random responses indicate that the total-test fit statistic identified 64% of the guessing cases (Table 4). In all, 79% of the guessing cases had one or more fit statistics above the critical level, while only 11% of these cases had fit statistics less than 1.0.

The choice of the critical fit statistic value of 1.5 is worth noting in light of the assumption that the fit statistic had a mean of zero and a standard deviation of 1. This statistic seems to be sensitive to test length. Since we are

dealing with fairly short tests and subgroups, some modification in the critical value of the fit statistic seemed necessary. For short tests it is not uncommon to find individuals with unacceptable response patterns in the 1.2 to 1.8 range. Thus, experience indicates the critical value should be set near 1.5.

Table 5 shows the frequency of misfitting individuals on each of the four misfit statistics. The frequency of misfit on the Elementary Algebra test is high. One third of the students who took this test did not fit the model. It would raise serious consequences in any testing situation if one third of the scores could not be considered valid. The frequency of misfit for the Logical Relationships test and the other three objective tests of like difficulty were similar. The number of misfitting individuals on each test was near 25%. It would be erroneous to attribute the incidence of misfit on the Elementary Algebra test solely to the difficulty of the items, but there seems to be some reason to suspect that there is a relationship between the average item difficulty of the test and the occurrence of misfit.

Tables 6 and 7 illustrate the identifiable patterns of misfit on the Elementary Algebra test. Approximately two-thirds of the individuals identified as misfitting fell into the range of one to nine correct answers, while only forty-five per cent of all the students taking the test fell into this range. It

manageable means of checking the quality of each student's results and of the ongoing utility of items used. This procedure can be incorporated into any automated scoring system. The procedure provides the means of checking all test results for misfit on a variety of user-defined dimensions. Very often the procedure can aid in the diagnosis of the type of misfit present and the development of appropriate action to produce a measure which fits our notion of measurement.

V. CONCLUSION

From what has been presented here, I hope we have demonstrated that measurement, at least as it is practiced in most standardized testing applications, is in fact incomplete if some effort is not made to insure the quality of the measure (ability estimate) which is reported for each individual. It should be clear that the traditional notion of the standard error of measurement is inappropriate for this task.

We have traced the development of the notion of appropriateness of an individual's response string from Thurstone's concept of psychophysical measurement through Rasch's formulation of the one parameter logistic model. We have discussed the unique qualities of the Rasch Model which enable us to gain a new perspective on the problem of analyzing response strings. We have demonstrated the utility of this approach in practical situations, presenting examples of the most common disruptions to measurement, and suggested means of eliminating the effects of these disruptions. Finally we have examined the frequency with which these disturbances occur in a college placement test.

From this, there are three conclusions which can be drawn. First, from a philosophical point of view, we must conclude that some measure of the quality of measurement is necessary if we are to use test data to make important decisions about individuals. Second, we must conclude that the

use of the properties of the Rasch measurement model provide a new and profitable method for exploring the concept of person disruptions to measurement. Finally, we hope that we have demonstrated that in real measurement applications the incidence of person disruptions can be so great as to invalidate up to 1/3 of the total number of scores reported. Obviously, it is impossible to generalize from the one situation studied here. But let us suppose that in most situations the frequency is only 1/3 as much as it was in this case. Would we or should we be satisfied with the scores reported when 1 out of every 10 scores could be invalidated by the presence of person disruptions?

REFERENCES

Andrich, D. Latent Trait Psychometric Theory in the Measurement and Evaluation of Essay-Writing Ability. Unpublished doctoral dissertation, University of Chicago, 1973.

Cronbach, L. J. Response Sets and Test Validity. Educational and Psychological Measurement, 1946, 6, 475-494.

Cronbach, L. J. Further Evidence on Response Sets and Test Design. Educational and Psychological Measurement, 1950, 10, 3-31.

Glaser, R. A Methodological Analysis of the Inconsistency of Responses to Test Items. Educational and Psychological Measurement, 1949, 9, 721-739.

Glaser, R. Multiple Operation Measurement. Psychological Review, 1950, 57, 241-253.

Glaser, R. The Application of the Concepts of Multiple Operation Measurement to the Response Patterns on Psychological Tests. Educational and Psychological Measurement, 1951, 11, 322-382.

Glaser, R. The Reliability of Inconsistency. Educational and Psychological Measurement, 1952, 12, 60-64.

Levine, M. V., & Rubin, D. B. Measuring the Appropriateness of Multiple-Choice Test Scores. Journal of Educational Statistics, 1979, 4, 269-290.

Lumsden, J. Test Theory. Annual Review of Psychology, 1976, 27, 251-280.

Lumsden, J. Person Reliability. Applied Psychological Measurement, 1979, 1, 477-482.

Lumsden, J. Tests Are Perfectly Reliable. British Journal Math. Statist. Psychol., 1978, 31, 19-26.

Lumsden, J. Variations on a Theme by Thurstone. Applied Psychological Measurement, 1980, 4, 1-7.

Mead, R. Assessing the Fit of Data to the Rasch Model. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1976.

Mead, R. PANAL: Person Analysis with the Rasch Model. University of Chicago, 1978.

Mead, R. J., & Kreines, D. C. Person Fit with the Dichotomous Rasch Model. Paper presented at the annual meeting of the American Educational Research Assocation, Boston, 1980.

Mead, R. J., & Wright, B. D. The Use of the Rasch Simple Logistic Model to Control the Effects of Random Guessing. Paper presented at the annual conference of the American Educational Research Association, New York, 1977.

Moiser, C. I. Psychophysics and Mental Test Theory: Fundamental Postulates and Elementary Theorems. Psychological Review, 1940, 47, 355-366.

Moiser, C. I. Psychophysics and Mental Test Theory. II. The Constant Process. Psychological Review, 1941, 48, 235-249.

Rasch, G. Probabilistic Models for Some Intelligence and Attainment Tests. Chicago: University of Chicago Press, 1980.

Sherif, M., & Cantril, H. The Psychology of Attitudes. Psychological Review, 1945, 52, 259-319; 1946, 53, 1-24.

Smith, R. M., & Mitchell, V. P. Improving the Predictive Ability of Placement Tests Using the Rasch Model for Scoring. Paper presented at the annual conference of The National Council on Measurement in Education, San Francisco, 1979.

Smith, R. M., & Mitchell, V. P. An Analysis of Rasch Person Fit Statistics in College Placement Tests. Paper presented at the annual conference of the National Council on Measurement in Education, Boston, 1980.

Thorndike, R. L. Personnel Selection: Test and Measurement Techniques. New York: Teachers College Press, 1949.

Thurstone, L. L. A Law of Comparative Judgment. Psychological Review, 1927, 34, 273-286.

Waller, M. I. Estimating Parameters in the Rasch Model: Removing the Effects of Random Guessing. Educational Testing Service, Princeton, N.J., March, 1976.

Weiss, D. J. The Stratified Adaptive Computerized Ability Test. (Research Report 8-73-3). Minneapolis: University of Minnesota, Dept. of Psychology, Psychometric Methods Program, 1973.

Wilson, E., & Hilferty, M. Distribution of Chi-Square. Proceedings of the National Academy of Sciences, 1931, 17, 684-688.

Wright, B. D. Solving Measurement Problems with the Rasch Model. Journal of Educational Measurement, 1977, 14, 97-116.

Wright, B. D. The Rasch Model for Test Construction and Person Measurement. Paper presented at the Fifth Annual Conference and Exhibition on Measurement and Evaluation, Los Angeles, 1978.

Wright, B. D., & Mead, R. J. BICAL: Calibrating Items and Scales with the Rasch Model. Research Memorandum Number 23A, University of Chicago, 1978.

Wright, B. D., & Panchapakesan, N. A. A Procedure for Sample Free Item Analysis. Educational and Psychological Measurement, 1969, 29, 23-48.

Wright, B. D., & Stone, M. H. Best Test Design. Chicago: MESA Press, 1979.

TABLE I

TEST INFORMATION

	Reading Comprehension	Sentence Structure	Logical Relationships	Computation	Elementary Algebra
Test Length	30 min.	30 min.	30 min.	35 min.	35 min.
No. of Items	40	35	50	30	30
State Wide (n=42950)					
Test Mean (STD Score)	64.5	64.2	64.7	63.9	65.4
STD. Dev. (STD. Score)	11.6	11.3	11.5	11.3	11.0
Test Mean (Raw Score)	30.6	24.7	38.7	21.2	14.0
Average p^+*	.765	.706	.774	.707	.467
College Sample N=822					
Test Mean (STD Score)	63.5	62.5	63.0	61.5	61.1
STD. Dev. (STD Score)	10.8	10.8	11.1	10.6	9.5
Test Mean (Raw Score)	29.9	23.7	37.6	19.8	11.0
Average p^+*	.748	.677	.752	.660	.367

*p^+ = The percentage of people responding to an item correctly.

TABLE 2

ITEM CLASSIFICATIONS

	Elementary Algebra	Logical Relations
SEQUENCE No. of Groups	3	2
No. of Items In Each Group	10, 10, 10	25, 25
SUB-SCORES No. of Groups	2	4
No. of Items In Each Group	18, 12	12, 13, 12, 13
DIFFICULTY No. of Groups	4	4
No. of Items In Each Group	6, 8, 11, 5	9, 18, 15, 8

TABLE 3

MEAN VALUE OF FIT STATISTICS

TEST	TOTAL TEST	SEQUENCE	SUBSCORES	DIFFICULTY	N
Random Responses	2.09	0.86	-0.08	1.63	47
Elementary Algebra	-0.06	0.25	0.38	0.17	760
Logical Relationships	-0.17	0.19	0.31	0.00	819

TABLE 4

PERCENT OF RANDOM RESPONSES PATTERNS
WITH CRITICAL FIT STATISTICS
(FIT $\geq$ 1.5)

	Number	% of Total (N=47)
Total Fit	30	64
Sequence	15	32
Subscores	4	9
Difficulty	29	62
One or More Critical Values	37	79
Fit < 1.0	5	11

TABLE 5

SUMMARY OF FIT STATISTICS

	Elementary Algebra	Logical Relationships
Total Sample	822	822
Perfect Scores	3	4
Zero Scores	59	2
Persons in Analysis	760	819
t* > 1.5	104 (14%)	89 (11%)
B1* > 1.5	94 (12%)	67 (8%)
B2* > 1.5	125 (16%)	82 (10%)
B3* > 1.5	79 (10%)	53 (6%)
t, B1, B2, B3 > 1.5	259 (34%)	204 (25%)
B3 > 1.5 and t < 1.5	21 (2%)	24 (3%)
B3 > 1.5 and t < 0	5 (1%)	4 (1%)

* t = Total Test Fit Statistic

B1 = Sequence Between Group Fit Statistic

B2 = Subscore Between Group Fit Statistic

B3 = Difficulty Between Group Fit Statistic

TABLE 6

SUMMARY OF TYPES OF MISFIT*

Elementary Algebra

	t>1.5	B1>1.5	B2>1.5	B3>1.5
Guessing	33	8	29	58
Start-up	1	5	-	7
Content Interaction	3	1	48	5
Speedness	1	-	-	2
Gave up	2	20	7	-
Missed 1 or 2 Very Easy Items	9	1	-	-
Worked 10 or more than Guessed	-	12	-	-

TABLE 7

SCORE DISTRIBUTION OF MISFITTING STUDENTS

Elementary Algebra

Score Range	t>1.5	B1>1.5	B2>1.5	B3>1.5
1 - 9	79	51	62	57
10 - 19	15	31	53	19
20 +	10	12	11	3

* Individuals with two critical misfit statistics were included under the catagory of greatest misfit.